



Calderhead, Ben (2012) *Differential geometric MCMC methods and applications*. PhD thesis.

<http://theses.gla.ac.uk/3258/>

Copyright and moral rights for this thesis are retained by the Author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Differential Geometric MCMC

Methods and Applications



Ben Calderhead

Department of Computing Science

University of Glasgow

A thesis submitted for the degree of

Doctor of Philosophy (PhD)

2011

Abstract

This thesis presents novel Markov chain Monte Carlo methodology that exploits the natural representation of a statistical model as a Riemannian manifold. The methods developed provide generalisations of the Metropolis-adjusted Langevin algorithm and the Hybrid Monte Carlo algorithm for Bayesian statistical inference, and resolve many shortcomings of existing Monte Carlo algorithms when sampling from target densities that may be high dimensional and exhibit strong correlation structure. The performance of these Riemannian manifold Markov chain Monte Carlo algorithms is rigorously assessed by performing Bayesian inference on logistic regression models, log-Gaussian Cox point process models, stochastic volatility models, and both parameter and model level inference of dynamical systems described by nonlinear differential equations.

Thesis Statement

This thesis is submitted in accordance with the regulations for the degree of Doctor of Philosophy at the University of Glasgow. Chapters 1 and 2, and the first half of chapter 3 contain known results and relevant background material presented from a statistical perspective. The remaining chapters are the author's original work, except where explicitly referenced. Parts of Chapters 3 and 4 have already been published jointly with my PhD supervisor Prof. Mark Girolami in the Journal of the Royal Statistical Society: Series B (with discussion) (77), as have parts of Chapters 5 and 6 in Neural Information Processing Systems and the Journal of the Royal Society Interface Focus (28, 29), respectively. No part of this thesis has previously been submitted for a degree at this or any other University.

Overview of Thesis

The quantification of uncertainty plays a vital role in almost every area of modern science and engineering, however it is only relatively recently that the wide availability of high performance computing has made the probabilistic analysis of more complex and realistic statistical models computationally tractable. Despite this there remain many models that are unapproachable using current methodology, particularly when employing computationally intensive Bayesian methodology, which provides a consistent framework for reasoning under uncertainty via the use of probability theory. Markov chain Monte Carlo (MCMC) methods can produce correlated samples from arbitrary posterior probability distributions, however complex statistical models often result in high dimensional, strongly correlated parameter spaces, for which standard MCMC approaches fare extremely badly.

In this work, we introduce ideas of Riemannian geometry as a means of creating novel and more efficient MCMC methods that work well in a wide variety of scenarios. As one motivation for this work, we consider a biological example of modelling circadian genetic networks using statistical models based on systems of nonlinear differential equations. Such models are widely applicable throughout the sciences and exhibit many inferential challenges to severely test new sampling methodology.

We begin in Chapter 1 by introducing some motivating examples and reviewing standard approaches to statistical inference over such models and the challenges they present. In Chapter 2 we review the use of Monte Carlo methods based on dynamical systems, as a means of improving sampling efficiency. We note that such dynamical methods are implicitly based on

a Euclidean space and that there is in fact much more geometric structure available, as was highlighted by Rao and Fisher in the 1940s.

In Chapter 3 we review ideas of Euclidean and Riemannian geometry and present generalisations of the Langevin and Hybrid Monte Carlo methods by defining them on a Riemannian manifold. We provide a thorough evaluation of these new sampling methodologies in Chapter 4, investigating a variety of challenging statistical models, including logistic regression, stochastic volatility, and log-Gaussian Cox point processes. Some of these models have unobserved high dimensional latent variables and structures that pose significant computational issues.

We focus in Chapter 5 on the task of performing statistical inference over systems of ordinary differential equations (ODEs). We describe the application of differential geometric MCMC methods to ODE models and compare the relative efficiency on a small biological example. We then consider an alternative fast approximate inference method for ODEs based on auxiliary Gaussian processes. Finally, we give a basic introduction and offer insight into the Bayesian analysis of dynamical models described by ODE systems using simple infection outbreak examples.

In Chapter 6 we tackle larger, biologically more realistic ODE models of biochemical systems. We investigate models describing circadian rhythms in the plant *Arabidopsis thaliana*, as well as a cell signalling network example. We demonstrate how differential geometric MCMC methods allow for efficient Bayesian inference and accurate estimates of marginal likelihoods, which ultimately allows us to consider the systematic comparison of competing model hypotheses to describe large and complex biological systems.

In Appendix A we offer a summary of the manifold MCMC methods developed in this thesis, along with detailed pseudocode and guidelines for their application.

Thesis Contributions

This thesis presents novel Markov chain Monte Carlo methodology that exploits the natural representation of a statistical model as a Riemannian manifold. The methods provide generalisations of the Metropolis-adjusted Langevin algorithm (MALA) and the Hybrid Monte Carlo (HMC) algorithm for Bayesian statistical inference. This methodology resolves many of the shortcomings of existing Markov chain Monte Carlo algorithms, particularly when sampling from high dimensional and strongly correlated probability distributions. In particular, the mathematical theory generalising MALA and HMC algorithms from a Euclidean space to a Riemannian manifold is derived, and the necessary concepts from Hamiltonian mechanics and differential geometry are presented from a statistical perspective.

A thorough numerical evaluation of these differential geometric MCMC methods is conducted, and the necessary equations are derived to apply these methods to a wide class of statistical models, including Bayesian logistic regression, stochastic volatility and log-Gaussian Cox point process models. The methods allow automated scaling by taking into account the local Riemannian structure at each point on the manifold of free parameters. This is especially clear in the log-Gaussian Cox example, where these methods require no tuning in the transient and stationary phases of the Markov chain; in contrast, standard approaches using a MALA method require different tuning parameters in each of these two regimes. These novel MCMC methods scale between $\mathcal{O}(n \log n)$ and $\mathcal{O}(n^3)$ depending on the statistical model of interest. Computationally more efficient versions are therefore also developed, which decrease the computation by assuming a locally constant metric on the Riemannian manifold. It is shown that these Riemannian manifold MCMC algorithms provide state-of-the-art sampling performance measured in terms of a time-normalised effective sample size.

Statistical models based on systems of differential equations are then considered and a novel sampling method is developed that employs Gaussian processes to approximate the model solutions. It is shown that this approximate inference method allows for parameter inference over ordinary and delay differential equations up to 2 orders of magnitude faster than alternative approaches. Finally, it is shown how differential geometric MCMC methods, combined with thermodynamic integration, allow for efficient Bayesian inference and accurate estimation of marginal likelihoods, which ultimately allows systematic comparison of competing model hypotheses that describe large and complex biological systems.

Acknowledgements

I would like to thank most of all my PhD supervisor, Professor Mark Girolami, for all the support and guidance he has offered for the duration of my academic career thus far. I would also like to thank Microsoft Research for providing me with the funding and academic freedom to carry out this research.

To my family - my brother, my parents and grandparents.

Contents

List of Figures	xii
List of Tables	xv
1 Introduction	1
1.1 Some Motivating Examples	3
1.2 Statistical Models based on Differential Equations	4
1.3 Computational Statistical Modelling	7
1.3.1 The Bayesian Approach	8
1.4 Monte Carlo Methods	9
1.4.1 Markov Chain Monte Carlo Methods	11
1.5 Conclusions	15
2 Dynamical MCMC Methods	16
2.1 Hamiltonian Dynamics	19
2.1.1 Hamilton's Equations	19
2.1.2 A Simple One Dimensional Example	21
2.1.3 Further Properties of Hamiltonian Systems	23
2.2 Molecular Dynamical Simulations	28
2.3 Methods of Numerical Integration	30
2.3.1 Euler's Method	30
2.3.2 Method of Splitting	31
2.3.3 Illustrative Example: Parameters of a Gaussian Distribution	35
2.3.4 Properties of Integration Schemes	36
2.4 Hamiltonian Monte Carlo	39
2.5 Metropolis-adjusted Langevin Algorithm	43

2.6	Conclusions	46
3	Riemannian Manifold MCMC Methods	48
3.1	Introduction	48
3.1.1	Why Consider a Riemannian Geometry?	49
3.1.2	A Quick Reminder of Euclidean Geometry	51
3.2	An Introduction to Riemannian Geometry	53
3.2.1	Differentiable Manifolds	54
3.2.2	Tangent Spaces	55
3.2.3	Metric Tensors	56
3.2.4	Riemannian Manifolds from Statistical Models	58
3.2.5	Choosing a Metric	62
3.2.6	Connecting Tangent Spaces	66
3.3	Riemannian Manifold MALA	71
3.3.1	Illustrative Example: Parameters of a Gaussian Distribution . .	76
3.4	Riemannian Manifold Hamiltonian Monte Carlo	76
3.5	Population Manifold Methods	80
3.6	Conclusions	81
4	An Evaluation of Manifold MCMC Methods	85
4.1	Methods of Comparison	85
4.1.1	Metropolis-Hastings	86
4.1.2	Metropolis Adjusted Langevin Algorithm	86
4.1.3	Hamiltonian Monte Carlo	87
4.1.4	Manifold Methods	87
4.2	Bayesian Logistic Regression	87
4.2.1	Experimental Results for Bayesian Logistic Regression	89
4.2.1.1	Auxiliary Variable Gibbs Sampler	90
4.2.1.2	Iterated Weighted Least Squares	90
4.2.2	Comparison of MCMC Methods	90
4.2.3	Comparison of mMALA and RMHMC Variants	92
4.3	Stochastic Volatility Model	97
4.3.1	mMALA and RMHMC for SVM Parameters	98
4.3.2	mMALA and RMHMC for SVM Latent Volatilities	99

4.3.3	Experimental Results for Stochastic Volatility Model	100
4.4	Log Gaussian Cox Model	101
4.4.1	Experimental Results for Log-Gaussian Cox Processes	106
4.5	Conclusions	111
5	Statistical Inference over Dynamical Systems	113
5.1	Manifold Sampling for ODE Models	116
5.1.1	First Order Sensitivities	117
5.1.2	Second Order Sensitivities	118
5.1.3	Fitzhugh Nagumo Model	119
5.1.4	Experimental Results	120
5.2	Gaussian Processes for Approximate ODE Inference	122
5.2.1	Overview	124
5.2.2	Introduction to Gaussian Processes	125
5.2.3	Auxiliary Gaussian Processes on State Variables	127
5.2.4	Sampling Schemes for Fully Observed Systems	128
5.2.5	Extension to Partially Observed Systems	131
5.2.6	Example 1 - Nonlinear Ordinary Differential Equations	131
5.2.7	Example 2 - Nonlinear Delay Differential Equations	134
5.2.8	Example 3 - The p53 Gene Regulatory Network with Unobserved Species	136
5.2.9	Discussion	137
5.3	Disease Outbreak Models	138
5.3.1	A Simple Infection Model	140
5.3.2	A More Realistic Infection Model	147
5.3.3	Model Selection	151
5.3.4	The Optimal Time to Escape	154
5.3.5	Discussion	156
5.4	Conclusions	156
6	Modelling Biochemical Dynamics	158
6.1	Mathematical Modelling	158
6.1.1	Circadian Rhythms in Arabidopsis Thaliana	162
6.1.2	Cell Signalling Networks	165

6.1.3	Parameter Identifiability	167
6.2	Implementation	168
6.2.1	The Choice of Priors and Ill-Conditioned Metric Tensors	170
6.3	Circadian Model Results	171
6.3.1	Linear Parameters	171
6.3.2	Transcription Parameters	175
6.3.3	Michaelis-Menten Parameters	175
6.3.4	Inference with Full and Partial Observations	179
6.3.5	Estimating Marginal Likelihoods for Model Ranking	180
6.4	Cell Signalling Model Results	183
6.5	Conclusions	188
7	Discussion	190
7.1	Conclusions	190
7.2	Future Work and Extensions	192
A	Manifold MCMC Recipes	194
A.1	Simplified Manifold MALA (SmMALA)	194
A.2	Manifold MALA (mMALA)	195
A.3	Riemannian Manifold Hamiltonian Monte Carlo (RMHMC)	195
A.4	Fixed Metric RMHMC	197
	Bibliography	200

List of Figures

1.1	A posterior distribution exhibiting strong correlation structure.	6
2.1	Comparison of integration schemes for Hamilton's Equations	36
3.1	Slow convergence of a Markov chain using Hamiltonian Monte Carlo . .	51
3.2	Representation of the tangent space on a sphere	56
3.3	Representation of a connection in a Riemannian manifold	68
3.4	Comparison of MALA, mMALA and simplified mMALA samplers using a simple Gaussian model	77
3.5	Example of tempered distributions	82
4.1	Autocorrelation of samples for Bayesian logistic regression with the Heart dataset	95
4.2	Posterior marginal distributions of the stochastic volatility model hyper- parameters	102
4.3	Comparison of sampling the latent volatilities in a stochastic volatility model using HMC and RMHMC	102
4.4	Close up of sampling the hyperparameters in a stochastic volatility model using HMC and RMHMC	103
4.5	Trace plots of the log joint probability of the log-Gaussian Cox model .	107
4.6	Posterior latent fields of the log-Gaussian Cox model	108
4.7	Kernel density estimates of the hyperparameters of the log-Gaussian Cox model	110
5.1	Fitzhugh Nagumo ODE model output.	120

5.2	Graphical models representing different approaches to inference over differential equation systems.	129
5.3	Summary statistics for time taken to perform ODE inference using a variety of methods.	133
5.4	Summary statistics for time taken to perform DDE inference using a variety of methods	135
5.5	The predicted output of the p53 gene using data from (12) and the accelerated GP inference method for the linear model	136
5.6	The predicted output of the p53 gene using data from (12) and the accelerated GP inference method for the nonlinear model	137
5.7	Example output from a simple infection model	141
5.8	Example output from a simple infection model	142
5.9	Varying numbers of data points generated from a simple infection ODE model	144
5.10	Posteriors inferred from a simple infection ODE model with a varying number of data points	144
5.11	Data points generated from a simple infection ODE model with varying noise	145
5.12	Posterior distributions inferred from simple infection ODE model with varying noise	146
5.13	Examining the effect of the prior on posterior inference over a simple infection ODE model	146
5.14	Examining the effect of the number of data points on posterior output from the complex infection ODE model	148
5.15	Posterior inference over the initial conditions of a complex infection ODE model	149
5.16	Prediction of future infection levels from a complex infection ODE model	150
5.17	Comparison of posterior outputs from two plausible infection models with low noise levels	153
5.18	Comparison of posterior outputs from two plausible infection models with medium noise levels	153
5.19	Comparison of posterior outputs from two plausible infection models with high noise levels	155

LIST OF FIGURES

5.20	Model predictions of disease outbreak in a small town	155
6.1	A representation of the model developed in (128), which we employ to model the main circadian oscillator network in <i>Arabidopsis thaliana</i> . .	166
6.2	Comparison of posterior samples of the linear parameters in the circadian model obtained using a variety of MCMC sampling algorithms	172
6.3	Pairwise scatter plots and density estimates of posterior samples of the linear parameters in the circadian model	174
6.4	Posterior distributions over the linear parameters of the circadian model using different priors	176
6.5	Pairwise scatter plots and density estimates of posterior samples of the Hill parameters in the circadian model	177
6.6	Comparison of sampling Michaelis-Menten parameters, with and without a bounded eigenvalue strategy	178
6.7	Scatter plots and density estimates of the pairs of Michaelis-Menten parameters from the circadian model	179
6.8	Population-based MCMC is extremely useful for ensuring convergence to the correct mode (27)	181
6.9	Comparison of posterior outputs from the circadian model with unobserved species	182
6.10	Comparison of posterior outputs from two fully observed circadian models	184
6.11	Posterior model predictions for the cell signalling model with additional observation model	185
6.12	Scatter plots and density estimates of the posterior distribution of a cell signalling model	186
6.13	Example path taken during the burn-in phase of a Markov chain exploring the posterior distribution of the cell signalling model	187

List of Tables

4.1	Summary of datasets for Bayesian logistic regression	89
4.2	RMHMC with generalised leapfrog integration scheme - investigating the effect of parameter settings on sampling efficiency with German Credit dataset	91
4.3	Comparison of sampling methods with Australian Credit dataset, $D = 14$, $N = 690$, 15 regression coefficients	92
4.4	Comparison of sampling methods with German Credit dataset, $D = 24$, $N = 1000$, 25 regression coefficients	93
4.5	Comparison of sampling methods with Pima Indian dataset, $D = 7$, $N = 532$, 8 regression coefficients	93
4.6	Comparison of sampling methods with Heart dataset, $D = 13$, $N = 270$, 14 regression coefficients	94
4.7	Comparison of sampling methods with Ripley dataset, $D = 2$, $N = 250$, 7 regression coefficients	94
4.8	Comparison of sampling the parameters β , σ and ϕ after 20,000 posterior samples averaged over 10 runs with 2000 simulated observations, $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$	100
4.9	Comparison of sampling the latent volatilities after 20,000 posterior samples averaged over 10 runs with 2000 simulated observations, $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$	100

4.10	Comparison of sampling methods for the latent variables of a log-Gaussian Cox Process. We note that MALA requires the use of a Cholesky decomposition at every iteration, which scales as $\mathcal{O}(n^3)$. The manifold methods, mMALA and RMHMC, require the inverse of the metric tensor to be calculated, which also scales as $\mathcal{O}(n^3)$, however this need only be calculated once as the metric is constant for the latent variables of this model.	110
5.1	Summary of results for the Fitzhugh Nagumo model with 10 runs of the parameter sampling scheme and 5000 posterior samples	122
5.2	Summary statistics for each of the inferred parameters of the Fitzhugh Nagumo model	132
5.3	Summary statistics for each of the inferred parameters of the Monk model	135
5.4	Interpretation of Bayes Factor	152
5.5	Summary of estimated marginal likelihoods for each infection model . .	154
6.1	Marginal likelihood estimates for each of the sets of synthetic observations for the circadian ODE model with negative feedback loop	183
6.2	Marginal likelihood estimates for each of the sets of synthetic observations for the alternative circadian ODE model based on a positive feedback loop	183

1

Introduction

Scientific models have been used in the biological natural sciences as far back as the 1800s when Darwin and his cousin Francis Galton attempted to explain the observed variability in the heights of self- and cross-fertilised Maize plants (47). It seemed to Darwin that cross-fertilised plants were generally taller than self-fertilised plants and he sought to prove this mathematically with the help of Galton, as Darwin himself was not terribly proficient at mathematics. They proposed a very simple parametric model taking into account just the systematic variation (due to the method of fertilisation) and the random variability (due to all other factors, not explicitly defined), and asked the question, was there a statistically significant difference in systematic variability between the two types of fertilisation?

It is interesting to see, even in this early stage of the scientific revolution of the 19th century, the importance and indeed presence of multidisciplinary collaboration that is nowadays beginning to define progress in the natural sciences and that will without a doubt play an even greater role in the future. The approach of devising a mathematical model to gain insight into the inner workings of biology is commonly employed today, albeit on a more detailed level. Nowadays of course, we have the computational advantage such that we need no longer restrict ourselves to mathematical descriptions that are analytically tractable.

Throughout this thesis we follow in the spirit of Darwin and Galton by seeking to develop methodology that allows us to reveal the structure and underlying mechanisms that operate in plants by enlisting the power of statistical reasoning. This thesis was initiated by a motivating example regarding the molecular genesis of circadian rhythms.

Such daily rhythms in plants have been studied at varying levels of detail for hundreds of years. Indeed even in the 1700s, the botanist Carl von Linne was able to construct a clock comprising plants and flowers with differing flowering times emerging from these underlying daily rhythms (124). As Darwin and others of his time were beginning to realise, characterising statistical variation of measured observations is an essential step towards explaining the underlying mechanisms that drive their behaviour; indeed this is the case not just in biology but all areas of the natural and physical sciences. This realisation heralded the start of a period of particular enlightenment in statistics (189).

Measurement is at the heart of any scientific procedure and we need a consistent and rigorous method of incorporating new measurements within a statistical model in order to update our current knowledge regarding the model parameters, as well as the model itself. The 1930s brought Kolmogorov's axioms for probability theory (111), and shortly thereafter Cox's derivation (41) making use of reasonable assumptions and desiderata. With this came a mathematically self-consistent method of incorporating new information from experimental data into currently held knowledge or beliefs encoded in the language of probability theory. Jaynes (97), in particular, argues eloquently that the Bayesian approach based on such an axiomatically derived system of probability is the only sensible way of proceeding; all other methods simply aim to approximate this gold standard.

Within this probabilistic Bayesian framework, we can imagine a sequence of analyses slowly but steadily converging on an ever more accurate mathematical description of the natural world as we measure it around us. Like a pixelated image that gradually comes into focus at higher and higher resolution, so our understanding of the world we observe increases as our measuring techniques become more and more accurate. Often in research, as in nature (and indeed in everyday life), one observes but the final outcome of a long and complicated process. It is easy to forget and often difficult to imagine the sequence of events that led to this end result. As interesting and useful as this final outcome may be, it is often the *process* that offers the greater insight. In this thesis, I shall therefore endeavour to convey not only the main ideas but also the process by which these ideas came about, in the hope of providing more insight and stimulating further discussion and research of the topics presented.

Cross disciplinary is becoming increasingly important, with statistical science underpinning a great many endeavours in a multitude of seemingly disparate research

areas, from neuroscience (1) to astrophysics (176), in a way that would not be possible were it not for recent advances in computing science. Just beneath the surface of these ventures we see the common task of reasoning with uncertain data and hypothesised mathematical models. Indeed, in the recent report *Towards 2020 Science* published by Microsoft Research (60), the authors note that An important development in science is occurring at the intersection of computer science and the sciences that has the potential to have a profound impact on science. The fact that Microsoft is funding important research into the natural sciences, and has funded the research contained in this thesis, highlights the significant impact computing science is poised to deliver.

This thesis focuses on methodological developments in computational statistics that have arisen through the consideration of a specific problem; describing and predicting the behaviour of circadian rhythms at a biomolecular level. We will revisit and draw on many important mathematical and conceptual developments, stretching back almost 400 years, from Newton’s description of gravity through classical mechanics, Bayes’ theory of inverse probability, Hamilton’s reformulation of classical mechanics, Riemann’s framework for describing curved geometries, Kolmogorov and Cox’s axiomatic derivations of probability theory, which lay the groundwork for modern Bayesian statistics, through to the Monte Carlo methods developed by the physicists Metropolis and Hastings, not to mention the countless number of smaller advances that helped these ideas reach their current states.

In short, we begin by looking for a particular solution to a specific problem, and finish by developing a very general methodology with a wide variety of applications.

1.1 Some Motivating Examples

We will develop and test novel methodology based on a number of statistical models that we believe characterise the main challenges encountered when reasoning under uncertainty using Bayesian inference. One such motivating example involves the modelling of plants. We are not interested in the variation of height that Darwin first investigated, but rather the circadian rhythms that play a central role in regulating their physiology.

Models based on differential equations can be used not only to describe circadian rhythms, but also many other biological processes at a molecular level, and these models

are generally characterised by large numbers of parameters, strong correlation structures and likelihoods that are computationally expensive to compute. Many other commonly employed statistical models share these properties; we shall see other examples in Chapter 4 that also have high dimensionality and whose structures pose significant computational problems, such as logistic regression models, stochastic volatility models, and models based on log-Gaussian Cox processes.

We shall use the example of differential equations as a starting point and return to it again in Chapters 5 and 6; we examine the specific modelling issues that arise in this context and use these as a motivation for developing new methodology that is relevant to a wider class of problems that share its characteristics.

1.2 Statistical Models based on Differential Equations

As our knowledge of a complex process or system increases, so generally does the size and sophistication of the statistical model required to accurately paint a mathematical picture of it. With this increase in model complexity comes an increased complexity of inference; we move from just a few model parameters to tens or perhaps hundreds, and we introduce more involved calculations often based on highly nonlinear dynamics. One natural approach to modelling such dynamics is to use differential equations to encode a hypothesis about the nature of interactions in a system (203).

A dynamical system can be defined by a system of ordinary differential equations (ODEs) that describe a deterministic functional relationship between the each of the n process states $\mathbf{x}(t) = [x_1(t) \dots x_N(t)]^T$ and their rate of change over time, such that

$$\dot{\mathbf{x}}(t) \equiv \frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \boldsymbol{\theta}, t) \quad (1.1)$$

We note that f is typically a nonlinear function of the the states at each time point, $\mathbf{x}(t)$, the models parameters $\boldsymbol{\theta} = [\theta_1 \dots \theta_D]^T$, as well as sometimes time itself, t . This defines a mechanistic model that can be used to describe the systematic components in a biological system. We can account for random variation (due to all other sources) through the use of some multivariate noise process $\boldsymbol{\epsilon}(t) = [\epsilon_1(t) \dots \epsilon_N(t)]^T$, which in turn can be defined in terms of some time-dependent function, whose exact form can depend on hyperparameters that may also be inferred from the experimental data.

1.2 Statistical Models based on Differential Equations

We therefore construct our statistical model and, in a similar way to Darwin and Galton, attempt to explain our data $\mathbf{y}(t) = [y_1(t) \dots y_N(t)]^T$ in terms of systematic and random components that are described through our mechanistic model and noise process respectively, such that

$$\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t) \tag{1.2}$$

We note at this point that there are also other formalisms that can be used for modelling dynamical systems. Whereas the inherently smooth solutions of ODEs may accurately describe the average behaviour of a large population of molecules, stochastic differential equations may be considered for systems with very small numbers of molecules, in which the randomness of individual molecules may play a much larger role in determining the overall behaviour of the system (204). Partial differential equations may also be more appropriate for modelling spatial behaviour of a system (129). The current availability of cellular transcript and proteomic assay techniques (149), which measure average concentrations in cell populations leads us to focus our attention on continuous deterministic ODE models for describing circadian rhythms, which are applicable not just in biology but also in a large number of other areas of science and engineering, e.g. (139). In addition, such models characterise many of the challenges associated with much wider classes of statistical models, and so we hope that any methodological advances we make may also be useful in other settings.

Although our approach to modelling is fundamentally the same as that employed by Darwin and Galton, we notice that there are several important differences. Firstly, we are modelling at a much finer granularity than the simple statistical model used to explain the measured height of plants. Modern science is now in a position to observe and measure changes all the way down to the molecular and genetic level of a biological system, and it is therefore this level of detail that we may endeavour to explain by building mathematical models. Secondly, there generally exist no analytic solutions for the large parameterised systems of ODEs that we can use to describe this complex biology. We must therefore employ computationally intensive numerical methods to solve our equations, often tens of thousands of times for different parameter combinations, before we find solutions that begin to describe our data and unlock insights into the underlying biology. Finally, the complex dynamics of nonlinear ODEs

1.2 Statistical Models based on Differential Equations

result in strong correlations between parameters, indeed sensitivity analysis (175) is often employed to determine the effect of changing parameter values, individually or in combination, on the output of the model. This correlation structure can drastically hamper efforts to optimise or sample parameters. An example of a posterior distribution induced by a relatively simple system of ODEs (26, 27) is shown in Figure 1.1, in which the oscillatory dynamics of the model result in ripples in the probability mass with multiple local maxima.

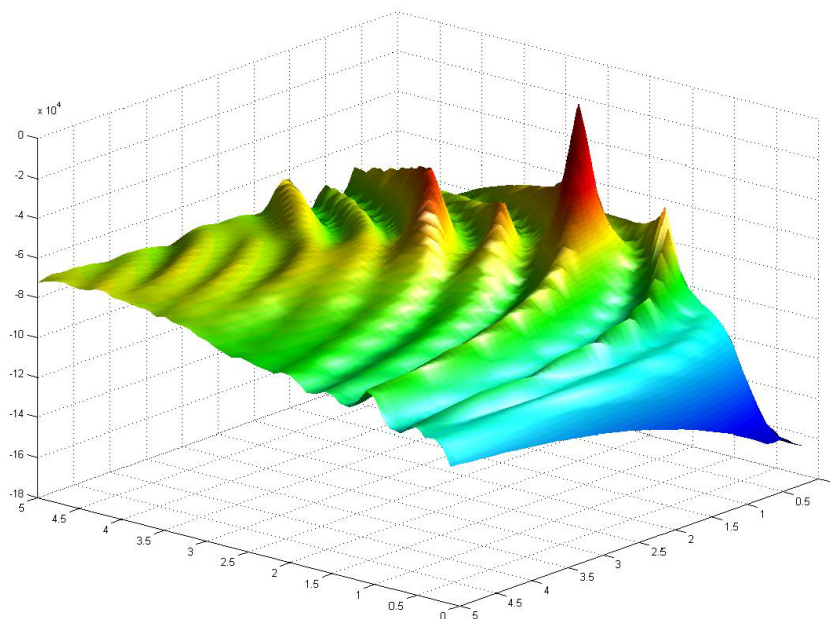


Figure 1.1: A posterior distribution exhibiting strong correlation structure. -

This plot demonstrates the complexity of a small system of nonlinear ordinary differential equations (26, 27). The ripples are produced by the model output moving in and out of phase with the oscillatory data for different parameter combinations. Such probability distributions are particularly challenging to sample from. The equations for this statistical model follow as $dx/dt = k_1/(36 + k_2y) - k_3$ and $dy/dt = k_4x - k_5$, with parameters $k_1 = 72$, $k_2 = 1$, $k_3 = 2$, $k_4 = 1$, $k_5 = 1$, and initial conditions $x(0) = 7$, $y(0) = -10$. 120 data points were simulated between $t = 0$ and $t = 60$, in steps of 0.5, and Gaussian noise was added with variance $\sigma^2 = 0.5$. This posterior plot was then calculated conditionally over the parameters k_3 and k_4 between 0 and 5.

Advances in scientific understanding are generally driven by the computational power and statistical techniques available to us. The use of ODEs allows greater complexity to be modelled and simulated as computer experiments. This approach pulls

theory and experimental work closer together; indeed predictive computer simulations can be used in place of physical experiments that are expensive or simply infeasible. In cases where parameters relate directly to rate constants and define physical relationships between interacting species, one may simply set the relevant parameters to simulate a particular physical situation and this can be a powerful means of furthering understanding, however we must take care when choosing the framework we wish to employ when reasoning and drawing conclusions from such models. Any inferences we make must be robust in the sense that they build upon our current knowledge in a consistent manner and in turn may be built upon in the future, when more data becomes available.

1.3 Computational Statistical Modelling

Decisions regarding our choice of modelling approach are often strongly influenced by the statistical and computational tools we have at our disposal. Before computing became widely available, more naive approaches to modelling complex systems were commonplace. Investigations often focused on the properties of individual models, and predictions were made using a single set of optimised parameters. This pragmatic approach was no doubt due to the limited computation that was possible given the tools of the time. Optimisation of a function is generally much faster than sampling, particularly for well-behaved, smooth functions such as those considered in this thesis; intuitively, obtaining a single solution is quicker than characterising all plausible solutions. Indeed it is perhaps no surprise that the rejuvenated interest in Bayesian methodology coincided with the arrival of high-speed computing. With increased computer power the focus is no longer on cleverly designing tractable approximate models, but rather on designing general, practical *methodology* that may be efficiently applied to a wide and general class of more accurate and complex statistical models.

Optimisation techniques are well suited for situations in which there is very little or no uncertainty in the system of interest, however applying it to situations where there is often large uncertainty, both in the data and the proposed model, presents a number of problems. Many models are unidentifiable such that there is no single most probable parameter value, but rather a collection of them. If we then wish to make predictions, it is useful to have an idea of how well we can trust the answers we

obtain; by formulating our answer in terms of a probability distribution we can assess the variance and indeed full covariance structure of the predicted parameter values, and have a built-in sensitivity analysis giving us information about the dependencies between model parameters.

The greater the complexity of our statistical models, the more risk there is of overfitting (50); given any data set we can design a model complex enough to describe all its features, however there is the danger of the statistical model describing the noise in the data instead of the true underlying relationship. The use of a more sophisticated system of reasoning allows us to guard against this. We would wish to be able to compare hypotheses and make predictions based on a deeper level of inference, in which uncertainties in the model are also taken into account. It is most appropriate to use a system of inductive reasoning when uncertainty is present, and in most areas of science it is exactly this kind of reasoning that we require.

Finally, we note that the probability distributions induced by more complex statistical models are sometimes multimodal, with local maxima and ridges that must be explored in order to find the globally maximum values. Such scenarios provide significant challenges to both optimisation and sampling methodologies, and it is worth noticing that the Markov chain Monte Carlo methods used in exploring the resulting probability distributions can be considered optimisation methods with an added detailed balance constraint, and so there is a natural connection between these two fields, although they have slightly different aims and applications.

1.3.1 The Bayesian Approach

Progress in the natural and physical sciences comes about by a process of developing hypotheses, running experiments, then refining the hypotheses based on observations. Descriptions of natural phenomena are often characterised by uncertainty, both in the data collected and subsequently in the models used to make sense of the data. In contrast to the deductive reasoning of Boolean logic (20), which defines an algebra of two values with statements logically being either true or false, Bayesian statistics allows for measures of uncertainty in a system to be propagated in a mathematically consistent manner. Herein lies the usefulness and wide applicability of Bayesian methods; by embedding inference within the mathematically sound framework of axiomatic *probability*, we avoid the need for ad hoc statistical constructions that may lead to inconsistencies

or even absurd results in extreme cases (97). Bayesian inference provides a consistent and rational framework for making sense of the world around us, letting us explicitly state our assumptions and update our current knowledge in light of newly acquired data.

Probability theory has been around since the 18th century (16, 117) as a means of making inferences in light of incomplete information. The axiomatic formulation of probability theory by Kolmogorov (111) together with a derivation by Cox (41) from a set of postulates that satisfy the desirable properties we would wish to have in a system of reasoning, have made Bayesian methods arguably the preferred method for inductive inference. Recent contributions by Knuth and Skilling (109) appear to add further support for the use of Bayesian probability; based on symmetry assumptions, they show that one is led to the probability calculus as the only logical and consistent calculus for reasoning under uncertainty.

Bayes theorem is simply an expression based on conditional probability and it states the conditional probability of an event A given an event B in terms of the probability of A , and the probability of B given A . In the context of a statistical model, the posterior distribution of the model parameters, $\boldsymbol{\theta} = [\theta_1 \dots \theta_D]^T$, given the data, $\mathbf{y} = [y_1 \dots y_N]^T$, is proportional to the prior distribution of the parameters multiplied by the likelihood of the data given the parameters.

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \quad (1.3)$$

Here the marginal likelihood in the denominator normalises the posterior density, such that it integrates to one and is a correctly defined probability distribution.

1.4 Monte Carlo Methods

For the purpose of making predictions, we often want to calculate expectations of a function with respect to the posterior distribution

$$\mu_f = E_{p(\boldsymbol{\theta}|\mathbf{y})}(f(\boldsymbol{\theta})) = \int f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \quad (1.4)$$

Since calculating an expectation is essentially just the same task as evaluating an integral, we could use quadrature methods and other numerical integration schemes. While these may offer great accuracy in lower dimensions, they scale extremely badly with the dimensionality of the problem and are therefore of little use when considering statistical models with large numbers of parameters. The Monte Carlo method (134) is a very useful approach for estimating such challenging integrals in high dimensions. It was developed in a US government laboratory at Los Alamos in the late 1940s for the purpose of modelling the random behaviour of subatomic particles in atomic bombs, and was no doubt inspired by previous attempts at estimating probabilities based on random simulation, such as the classic example of Buffon's needle during the 18th century (2). The Monte Carlo method has since found much wider applicability, particularly in Bayesian statistics, which is rather dependent on calculating high dimensional integrals. Given N samples from $p(\boldsymbol{\theta}|\mathbf{y})$ such integrals can be approximated using the Monte Carlo estimator (167),

$$\hat{\mu}_f = \frac{1}{N} \sum_{i=1}^N f(\boldsymbol{\theta}) \quad (1.5)$$

This estimator is unbiased and converges almost surely to the true integral. Assuming the variance of f , $\hat{\sigma}_f^2 = E_{p(\boldsymbol{\theta}|\mathbf{y})}[f^2(\boldsymbol{\theta})] - \mu(f)^2$ is finite, we can also obtain the variance of the estimator,

$$\text{var}(\hat{\mu}_f) = \text{var}\left(\frac{1}{N} \sum_{i=1}^N f(\boldsymbol{\theta})\right) = \frac{\hat{\sigma}_f^2}{N} \quad (1.6)$$

The advantage of this estimator is that the rate of convergence is independent of the dimensionality of $\boldsymbol{\theta}$, assuming the samples are independent of one another. It can however be tricky to draw independent samples from high dimensional and strongly correlated probability distributions. The Monte Carlo estimator is our method of choice given its theoretical indifference to dimensionality and in the rest of this thesis we shall address the problem of how to draw independent, or least minimally correlated, samples from distributions of interest.

1.4.1 Markov Chain Monte Carlo Methods

The use of Markov chains to produce samples from an arbitrary probability distribution was first suggested in the 1950s by Metropolis et al. (133). This approach allows us to simulate a Markov chain such that its stationary distribution is in fact the target distribution we are interested in; in other words the samples obtained from our Markov chain are equivalent to correlated samples drawn from the target distribution, and this may be implemented even when the normalising constant is unknown. The amount of correlation impacts the variance of the Monte Carlo estimator, and we therefore wish to use a Markov chain Monte Carlo (MCMC) method that reduces this as much as possible.

We begin by defining a transition density that dictates how the chain explores the parameter space. The aim is to find a transition density that proposes new minimally correlated parameter values that are accepted with high probability. The Metropolis-Hastings algorithm proceeds as follows

Algorithm 1 Standard Metropolis-Hastings Algorithm

- 1: Given current state θ , draw proposed state θ^* from transition density $T(\theta^*|\theta)$
 - 2: Calculate the acceptance ratio $R(\theta^*|\theta) = \min \left[1, \frac{p(\theta^*)T(\theta|\theta^*)}{p(\theta)T(\theta^*|\theta)} \right]$
 - 3: Draw $U \sim \text{Uniform}[0, 1]$
 - 4: Let $\theta = \begin{cases} \theta^* & \text{if } U < R(\theta^*|\theta) \\ \theta & \text{otherwise} \end{cases}$
-

Thus the new set of parameters θ^* are accepted with probability R . We note that this form is due to the generalisation by Hastings (87) that allows T to be any normalised probability distribution, and in the original method by Metropolis the transition T was required to be symmetric. Indeed, it was Hastings who first described this algorithm in its more general form in terms of Markov chains sampling from an arbitrary target distribution $\pi(\theta)$; before this it was described purely in physical terms using the motivating example of statistical mechanics. Despite this publication, it was almost another 20 years until its utility to Bayesian statistics began to be fully realised (71). There are many excellent expositions on the Metropolis-Hastings algorithm (37, 74, 101, 125, 167), and this is no doubt linked to the fact that it is considered one of the most important of all Monte Carlo algorithms (18).

We can show that such a Markov chain does indeed converge to the required stationary distribution by considering the following transition function, which we denote $A(\boldsymbol{\theta}^*|\boldsymbol{\theta})$. This is the total probability of firstly a proposed point being sampled from T , and secondly this proposed point actually being accepted with probability R ,

$$A(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = T(\boldsymbol{\theta}^*|\boldsymbol{\theta})R(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \quad (1.7)$$

A Markov chain will converge (167) if

$$\int p(\boldsymbol{\theta})A(\boldsymbol{\theta}^*|\boldsymbol{\theta})d\boldsymbol{\theta} = p(\boldsymbol{\theta}^*) \quad (1.8)$$

In other words the average probability of moving from any point in parameter space, denoted by $\boldsymbol{\theta}$, to a particular point $\boldsymbol{\theta}^*$ is equal to the probability of $\boldsymbol{\theta}^*$ itself. This must hold for all points $\boldsymbol{\theta}^*$.

A Markov chain will also converge under the following, more restrictive, condition known as detailed balance

$$p(\boldsymbol{\theta})A(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = p(\boldsymbol{\theta}^*)A(\boldsymbol{\theta}|\boldsymbol{\theta}^*) \quad (1.9)$$

and chains that satisfy this symmetry constraint are known as reversible, since the probability of moving from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$ is the same as moving from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}$, for all values of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. Assuming detailed balance, we can see straightforwardly that Equation 1.8 holds

$$\int p(\boldsymbol{\theta})A(\boldsymbol{\theta}^*|\boldsymbol{\theta})d\boldsymbol{\theta} = \int p(\boldsymbol{\theta}^*)A(\boldsymbol{\theta}|\boldsymbol{\theta}^*)d\boldsymbol{\theta} \quad (1.10)$$

$$= p(\boldsymbol{\theta}^*) \quad (1.11)$$

since A is a normalised probability distribution that by definition integrates to 1. In order to show that using the acceptance ratio R satisfies detailed balance, we again see straightforwardly that

$$p(\boldsymbol{\theta})A(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = p(\boldsymbol{\theta})T(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \min \left\{ 1, \frac{p(\boldsymbol{\theta}^*)T(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})T(\boldsymbol{\theta}^*|\boldsymbol{\theta})} \right\} \quad (1.12)$$

$$= \min \{ p(\boldsymbol{\theta})T(\boldsymbol{\theta}^*|\boldsymbol{\theta}), p(\boldsymbol{\theta}^*)T(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \} \quad (1.13)$$

$$= p(\boldsymbol{\theta}^*)A(\boldsymbol{\theta}|\boldsymbol{\theta}^*) \quad (1.14)$$

Alternative acceptance rules are explored in (13), however Peskun, a PhD student of Hastings, showed that the form given by Hastings was asymptotically optimal (154). There are also a couple of technical conditions on a Markov chain that are required for it to converge (167); a chain must be *irreducible* and *aperiodic*, which state that a chain has a non-zero probability of reaching any point in parameter space from any other point within a finite number of steps, and that the chain does not get stuck in any loops such that it repeatedly visits the same location with some fixed regularity.

Constructing a Markov chain that adequately samples from the target distribution however is not straightforward in practice due to the following three main issues.

Global Convergence Converging from a random starting position in parameter space to the true stationary distribution may be difficult, due to the possibility of multiple modes. We require sampling methods that can explore all modes globally with the correct frequency without becoming stuck in local maxima of negligible probability mass.

Local Mixing Once the Markov chain reaches a region of high probability mass, we want it to fully explore and ideally obtain near-uncorrelated samples. This is challenging for many types of differential equation based models, particularly those that are high dimensional and whose nonlinear dynamics induce very strong nonlinear correlation structures in the posterior distribution.

Computational Cost The likelihood of these models can be expensive to evaluate, since it involves approximately solving the system of ODEs with a numerical integration scheme for each set of proposed parameters. Although generally unavoidable, this cost can be minimised through efficient exploration of the parameter space, and measured in terms of effective sample size (ESS), normalised by the overall computational time (77).

In practice the proposal mechanism determines how efficiently a Markov chain can explore the space, and this becomes particularly important in high dimensional and strongly correlated parameter spaces, although even in low dimensions, statistical models based on nonlinear differential equations can induce complex, multimodal posterior densities (see Figure 1.1 as an example). For exploring multimodal posterior distributions, it is often useful to employ some form of tempered MCMC method. Similar to simulated annealing (108), the idea is to explore a collection of tempered distributions, in which the Markov chains are allowed to interact and swap positions with each other, whilst maintaining detailed balance, allowing for easier exploration of the target distribution. In addition, the samples drawn from the tempered distributions may be employed to obtain accurate estimates of the marginal likelihood and thus calculate Bayes factors for model comparison via thermodynamic integration (27, 67, 118). We may choose any MCMC method to sample within each tempered distribution and we shall consider a specific example in Chapter 5 that employs the manifold MCMC methodology we develop in Chapter 3.

A great amount of research in MCMC methodology has been conducted over the last 20 years. Alternative sampling approaches include slice sampling (143, 147) and nested sampling (182). Reversible Jump MCMC can be used for sampling between parameter spaces of varying dimension (79). There are also a number of approximate methods available, such as the recent INLA method (174), which are based on Laplace approximations and can often be much faster than MCMC sampling. One must bear in mind however that such approaches are only approximate and often it is difficult to quantify the errors associated with the answers obtained, as mentioned in the discussion section of (174). In addition such methods often make assumptions regarding conditional independence of parameters, and so information about parameter correlations in the posterior distribution can be lost.

In this thesis we concentrate on developing computationally efficient methods that retain all the information present in the posterior distribution and are exact in the sense that arbitrary accuracy may be obtained by collecting as many posterior samples as we require. Such an approach therefore acts as a gold standard when performing Bayesian inference over complex statistical models. In particular we focus on extending the dynamical MCMC methods introduced in Chapter 2, which are based on Langevin

diffusions and Hamiltonian mechanics, in order to make use of the intrinsic geometry of the probability distribution of interest.

1.5 Conclusions

Statistical models are being used to describe the natural world with increasing levels of sophistication, and the use of probability theory through Bayesian statistics allows us to naturally account for any sources of uncertainty, both in the measurements we make and in the models we employ. Useful quantities in Bayesian statistics may be estimated using Monte Carlo methods, and in particular MCMC methods may be used to draw samples from the complex probability distributions induced by our statistical models. The accuracy of the Monte Carlo method, although independent of the dimensionality of the problem, is determined not only by the number of samples available, but also by the magnitude of correlation within those samples, which may have a significant effect on the variance of any estimates. The task of drawing independent samples from arbitrary probability distributions is challenging due to the complex correlation structure that is often present.

In the context of modelling biological systems, this structure in the posterior distribution is closely related to the idea of sensitivity analysis, since changes in the model parameters not only have an effect on the model output, but also on the probability mass associated with the parameters. A particularly interesting and useful class of MCMC methods that we have not yet mentioned are those based on dynamical systems, which make use of 1st order sensitivities in order to propose better moves within an MCMC algorithm. The dynamical MCMC methods we review in Chapter 2 are intimately tied to the study of dynamical systems and are implicitly defined on the natural Euclidean geometry associated with the parameter space. In Chapter 3, we will expand these ideas further by considering higher order sensitivities and exploring the deeper links with Riemannian geometry, with the aim of developing MCMC methods that converge quicker to the stationary distribution and exhibit lower correlation in the final samples. This will ultimately allow us perform statistical analyses of a wide range of models, and in particular those based on differential equations, which can be used to describe complex biological processes.

2

Dynamical MCMC Methods

Shortly after the introduction of Monte Carlo methods in the 1950s, molecular dynamical simulations were introduced as a means of directly probing the physical properties of chemical systems by explicitly calculating the dynamics of individual molecules and their interactions with one another. Such dynamics are naturally described in terms of the rates of change of the interacting molecules and may therefore be conveniently described using differential equations. In particular, these molecular dynamics can be modelled as Hamiltonian systems, which are derived directly from Newton's 2nd law of motion, describing the acceleration of individual particles in terms of the forces acting upon them. The aim of molecular dynamics is to provide information regarding the *average* behaviour of the system under study. Since precise initial conditions for individual molecules are unknown it is hoped that by averaging their physical behaviour over time, the initial conditions become irrelevant; the overall properties of the system are then represented in terms of a stationary distribution that describes the probability of finding the system in any particular state.

In Bayesian statistics we are also interested in calculating averages over a range of states. In particular, we often wish to calculate high dimensional integrals that are written in terms of a particular function *averaged* over a probability distribution, which may be evaluated using a Monte Carlo estimator. The Markov chain Monte Carlo and Molecular Dynamics methods are closely linked through ergodic theorems, which state that under certain conditions the time average of a dynamical system converges to the same value as the space average of all the possible states of the system.

A Markov chain must be aperiodic, irreducible and positive recurrent in order to be ergodic, such that it will eventually visit all points of the space and that the time average will equal the space average. While the continuous-time Hamiltonian dynamics, upon which Molecular Dynamics are based, also have a time average that converges to the space average, their discretised approximate solutions do not. This is perhaps the main shortcoming of the Molecular Dynamics approach; often the Hamiltonian dynamics cannot be calculated analytically and the approximate samples are therefore not drawn from the true stationary distribution. The averages obtained are often assumed to be “accurate enough” to be of use describing the average properties of the underlying physical system, since in practice they are indeed often seen to be numerically very close to the true solutions, given a small enough step size.

These two computational approaches for determining the average properties of a physical system followed quite separate paths of development until the seminal paper of Duane et al. (57), in which the authors proposed a “hybrid” algorithm combining the two ideas; this was aptly named the Hybrid Monte Carlo method. Following more recent convention however, we shall refer to this approach using the more suitably descriptive title of the same acronym, Hamiltonian Monte Carlo (HMC), emphasising the fact that MCMC proposals are made based on the Hamiltonian dynamics of the underlying system, which are described in terms of a system of ordinary differential equations. The use of a Metropolis-Hastings acceptance step on the HMC trajectories corrects the error introduced by the discretisation of the continuous equations and ensures that samples are drawn from the true stationary distribution.

Molecular dynamical simulations can also be conveniently modelled using Langevin diffusions, which are described in terms of stochastic differential equations and whose time evolution of states also tends to a stationary distribution in the continuous-time case. Once again, however, discretisation of these Langevin dynamics causes time averages to converge to the wrong stationary distribution, and a Metropolis-Hastings acceptance step is needed to correct this problem.

For the purpose of efficiently obtaining low variance Monte Carlo estimates we wish to have a Markov chain that converges quickly to the stationary distribution, produces samples from the posterior distribution with low correlation, and has proposed moves accepted with high probability. The most basic form of Metropolis-Hastings algorithms employ random walks whereby the proposed steps are generated by a proposal density

that is independent of the target distribution. Often these take a particular parametric form, which can be tuned to obtain the desired acceptance rate, then fixed such that samples are drawn from the stationary distribution. Hamiltonian systems and Langevin diffusions exhibit many properties that make them attractive to use for developing efficient MCMC methods, and we shall explore these later in this chapter. Such approaches are often found to be more efficient than simple random-walk Metropolis algorithms, both in terms of sampling correlation and convergence, since they make use of additional geometric information from the target distribution in the form of first order derivatives. Intuitively one might expect that proposed steps using gradient information will make better moves, since they can follow the gradient to points of higher probability rather than relying on randomly chosen steps centred at the current point. Indeed this intuition appears to be true when we consider the theoretical optimal acceptance ratios; the optimal for gradient based sampling methods are higher than basic random walk Metropolis-Hastings (170).

Of course higher acceptance rates do not automatically mean more efficient sampling, since this could be achieved simply by making smaller proposed steps that are accepted more regularly but result in higher correlation between samples. Ultimately, the overall efficiency of any method must also be measured against any increases in the computational overhead of the sampler. For these dynamical methods we might then ask whether it is much more expensive to obtain the required geometric information? If it is more expensive, we might well be better running a simpler and computationally less expensive method for a larger number of iterations to achieve a similar accuracy of estimate. On the other hand, for more complex models there may well also be a large cost involved with evaluating the likelihood; in the statistical models considered in Chapters 5 and 6 for example calculating the likelihood involves solving a complex system of differential equations. An efficient method will therefore also be one that minimises the number of likelihood evaluations and takes larger steps in parameter space that are accepted with greater probability.

We note in passing that one advantage of employing a gradient-based MCMC method is that these gradient evaluations may usefully be reused. For example, certain variance reduction methods can make use of gradient information to achieve lower variance MCMC estimates (137), meaning that fewer samples are necessary to obtain the same level of accuracy as the standard MCMC estimator.

We now begin by reviewing some of the theory of Hamiltonian dynamics (Section 2.1), deriving Hamilton’s equations from Newton’s 2nd law of motion and investigating some of its most important properties, in particular its symplectic geometric nature, which is vital for developing valid MCMC schemes. We describe its use in molecular dynamical simulation via discretised computation (Section 2.2), and discuss the challenges this introduces. In Section 2.3 we have a look at methods of numerical integration and the geometric properties that characterise methods that are particularly suitable for accurately solving conserved systems, and we see that it is geometry that intimately ties ideas from Hamiltonian dynamics with those from probability theory. We then present the combination of Molecular Dynamics and Monte Carlo through the Hamiltonian (Hybrid) Monte Carlo method in Section 2.4, discussing the properties, advantages and difficulties of this approach. Finally, we review a stochastic approach to molecular modelling based on a Langevin diffusion (Section 2.5), showing that it corresponds to a special case of HMC.

2.1 Hamiltonian Dynamics

The Hamiltonian formalism has played a hugely influential role in computer simulation within the natural and physical sciences over the last half a century. Hamiltonian dynamics were first presented in 1834 by William Rowan Hamilton as a reformulation of classical mechanics based on Newton’s equations of motion (85). Given a closed system in which the total energy is conserved, Hamilton’s equations provide an alternative and equivalent way of describing how particles in the system deterministically evolve over time according to the laws of classical physics.

2.1.1 Hamilton’s Equations

Hamiltonian dynamics can be obtained as a reformulation of Newton’s 2nd law,

$$\mathbf{a} = \frac{d^2\boldsymbol{\theta}}{d\tau^2} = \mathbf{M}^{-1}\mathbf{f} \quad (2.1)$$

where \mathbf{a} is an acceleration vector, \mathbf{M} is a mass matrix, \mathbf{f} is a force vector, τ is time and $\boldsymbol{\theta}$ is a state or position vector. It is also equivalent to the Lagrangian formalism of classical mechanics (83), which imposes second order differential constraints on an

n -dimensional coordinate space. The Hamiltonian formalism however allows for easier solution of the system's evolution as it describes the same system in terms of only first order differential constraints. The price that must be paid for this reduction in differential order is that the coordinate space doubles in size; Hamiltonian systems are described in a $2n$ -dimensional space, with n values describing position and a further n values describing the associated momentum in each direction. In a statistical context the n position variables $\boldsymbol{\theta}$ correspond to the n parameter values of a statistical model.

The total energy of this Hamiltonian system is denoted by H and is composed of kinetic energy and potential energy, with the assumption that the total energy is conserved. Hamiltonians can therefore usefully be employed to describe complex dynamical systems that are devoid of friction. Interestingly such scenarios often occur on the very large scale and the very small scale, such as celestial mechanics and molecular dynamics. The standard form of the Hamiltonian is given as

$$H(\boldsymbol{\theta}, \mathbf{p}) = E_p(\boldsymbol{\theta}) + E_k(\boldsymbol{\theta}, \mathbf{p}) \quad (2.2)$$

where E_k is the kinetic energy, E_p is the potential energy, and H is a constant equal to the total energy in the system. If the kinetic energy is a function of momentum only, the Hamiltonian is termed separable, otherwise it is a non-separable Hamiltonian. These different types must be solved using different methods and induce differing underlying geometries. We shall return to the differences later in the chapter and for now consider only the simpler separable Hamiltonian. Hamilton's first order equations are defined as

$$\frac{d\boldsymbol{\theta}}{d\tau} = \frac{\partial H}{\partial \mathbf{p}}, \quad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial H}{\partial \boldsymbol{\theta}} \quad (2.3)$$

and these may be solved to obtain solutions for the evolution of the position and momentum variables of the Hamiltonian over time, τ . By differentiating Equation 2.2 with respect to time, a simple application of the chain rule shows that the rate of change of the Hamiltonian is zero, and therefore the total energy is indeed constant.

$$\frac{dH(\boldsymbol{\theta}, \mathbf{p})}{d\tau} = \frac{\partial H}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial \tau} + \frac{\partial H}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \tau} \quad (2.4)$$

$$= -\frac{\partial \mathbf{p}}{\partial \tau} \frac{\partial \boldsymbol{\theta}}{\partial \tau} + \frac{\partial \boldsymbol{\theta}}{\partial \tau} \frac{\partial \mathbf{p}}{\partial \tau} \quad (2.5)$$

$$= 0 \quad (2.6)$$

2.1.2 A Simple One Dimensional Example

We may gain some insight into what exactly these equations are describing by considering a one-dimensional example. The kinetic energy is defined in terms of mass and velocity as

$$E_k \equiv \frac{1}{2}mv^2 \quad (2.7)$$

$$= \frac{p^2}{2m} \quad (2.8)$$

which we can express in terms of the momentum p by using the expression $p = mv$. A one dimensional Hamiltonian may therefore be written as

$$H(\theta, p) = \frac{p^2}{2m} + E_p(\theta) \quad (2.9)$$

where for now we leave the potential energy simply as some function of the position coordinate. Let us consider first how the position coordinates change over time.

$$\frac{d\theta}{d\tau} = \frac{\partial H}{\partial p} = \frac{\partial}{\partial p} \left[\frac{p^2}{2m} \right] = \frac{p}{m} = v \quad (2.10)$$

We see from Hamilton's equations that as expected, the rate of change of position is simply the velocity with respect to time. It is more interesting to consider how the momentum changes over time,

$$\frac{dp}{d\tau} = \frac{d}{d\tau} [mv] = ma = f \quad (2.11)$$

and

$$\frac{dp}{d\tau} = -\frac{\partial H}{\partial \theta} = -\frac{\partial E_p(\theta)}{\partial \theta} \quad (2.12)$$

The change of momentum is simply equal to the Newtonian force on our imaginary particle, and via Hamilton's equations we see that this force is equal to the rate of loss of the potential energy with respect to the position coordinates θ . We can now verify for this one dimensional Hamiltonian example that the energy is indeed conserved. We proceed by once again differentiating the Hamiltonian with respect to time.

$$\frac{dH}{d\tau} = \frac{d}{d\tau} \left[\frac{1}{2}mv^2 \right] + \frac{d}{d\tau} [E_p(\theta)] \quad (2.13)$$

$$= \frac{1}{2}m \frac{d}{d\tau} [v^2] + \frac{dE_p(\theta)}{d\theta} \frac{d\theta}{d\tau} \quad (2.14)$$

$$= mv \frac{F}{m} + \frac{dE_p(\theta)}{d\theta} v \quad (2.15)$$

$$= Fv - Fv \quad (2.16)$$

$$= 0 \quad (2.17)$$

Hamilton's equations can be derived directly from Newton's second law, which states that an object's acceleration is directly proportional to the force applied to it and inversely proportional to its mass. We have already seen that this applied force is equal to the rate of the loss of potential energy with respect to the position,

$$\mathbf{f} = \mathbf{M}\mathbf{a} \quad (2.18)$$

$$= \mathbf{M} \frac{d^2 \boldsymbol{\theta}}{d\tau^2} \quad (2.19)$$

$$= \frac{d}{d\tau} [\mathbf{M}\mathbf{v}] = -\frac{dE_p(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \quad (2.20)$$

where we have written the mass in a more general form as a matrix. Since $\mathbf{p} = \mathbf{M}\mathbf{v}$, we have

$$\frac{d}{d\tau} [\mathbf{p}] = -\frac{dE_p(\boldsymbol{\theta})}{d\boldsymbol{\theta}}, \quad \frac{d}{d\tau} [\boldsymbol{\theta}] = \mathbf{M}^{-1} \mathbf{p} \quad (2.21)$$

where the second equation is simply equal to the velocity \mathbf{v} . There are of course many other ways in which Newton's second law could be rewritten in terms of two first order

differential equations, however they are conveniently written in this form with the observation that they may be derived from derivatives of the same underlying function. This Hamiltonian is therefore a function of position $\boldsymbol{\theta}$ and momentum \mathbf{p} ,

$$H(\boldsymbol{\theta}, \mathbf{p}) = E_p(\boldsymbol{\theta}) + \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} \quad (2.22)$$

Here the kinetic energy takes a quadratic form consisting of the mass and momentum terms, and we note in passing the similarity to the logarithm of an unnormalised Gaussian distribution. This allows us to give a statistical interpretation of this Hamiltonian, which we review in Section 2.4.

2.1.3 Further Properties of Hamiltonian Systems

It is clear that there is an antisymmetry present in Hamilton's equations (2.3), which we can make explicit by writing them in a more compact form. Let us concatenate the position and momentum vectors, such that $\mathbf{z} = [\boldsymbol{\theta}, \mathbf{p}]$ is a single vector. Hamilton's equations may then be written as

$$\frac{d}{d\tau}[\mathbf{z}] = \mathbf{J} \nabla_{\mathbf{z}} H(\mathbf{z}) \quad (2.23)$$

where \mathbf{J} is known as the structure matrix (122) and for this example has the form

$$\mathbf{J} = \begin{bmatrix} 0 & \mathbf{I}_n \\ -\mathbf{I}_n & 0 \end{bmatrix} \quad (2.24)$$

Other forms of Hamiltonian systems may be expressed in this standard form with a suitable choice of \mathbf{J} ; (122) gives an example of a Hamiltonian with an alternative \mathbf{J} that describes the motion of a charged particle in a magnetic field. For our current purposes however the form above is sufficient, and indeed induces a specific *symplectic* structure that will be a key feature of the Hamiltonian systems we are interested in. Symplecticness implies the existence of integral invariants (122), in particular the invariance of volume, and we shall algebraically define this shortly.

Hamiltonian systems exhibit many properties that we will see are particularly useful in proving that their dynamics may be used to form the basis of a valid MCMC

method. Some of these properties may be elucidated through the use of a flow map for a Hamiltonian. The flow map $(\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)) = \Phi_\tau(\boldsymbol{\theta}(0), \mathbf{p}(0))$ is a mapping from a set of coordinates and momentum values at time $\tau = 0$ to another set of coordinates and momentum values at some time τ . Such mappings are given by the deterministic solution of Hamilton's equations. As we have already seen, all exact solutions to Hamilton's equations preserve the initial total energy such that $H(\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)) = H(\boldsymbol{\theta}(0), \mathbf{p}(0))$. In addition, we shall show shortly that Hamiltonian dynamics are reversible and their solutions volume preserving, two very useful properties for generating proposals in an MCMC method.

Reversibility of Hamilton's equations follows straightforwardly from the standard theory of ordinary differential equations (9). Picard's theorem guarantees the existence and uniqueness of solutions to first order differential equations, given suitable initial conditions. As a result, the flow map $\Phi_\tau(\boldsymbol{\theta}(0), \mathbf{p}(0))$ is a bijection, whose invertibility guarantees the reversibility of the dynamics; indeed this inverse map is obtained simply by analytically integrating backwards in time.

Hamilton's equations are also volume preserving, and this property will be useful later for developing correct MCMC schemes based on Hamiltonian dynamics. Volume preservation is the geometric concept that when points in the phase space, $(\boldsymbol{\theta}(0), \mathbf{p}(0))$, undergo a transformation, $\Phi_\tau(\boldsymbol{\theta}(0), \mathbf{p}(0))$, the volume enclosed by the set of points does not change.

Volume preservation is actually a weaker property implied by the fact that Hamiltonian systems are *symplectic* (83, 122). This property is induced by the structure matrix \mathbf{J} described in Equation 2.24 and is defined in terms of the sum of areas of the parallelograms induced by pairs of vectors. We may see this property more clearly by first considering two 2-dimensional vectors. We note that since the Hamiltonian phase space consists of position and momentum, it will always be $2d$ -dimensional and can therefore be split up into d 2-dimensional vectors, each consisting of a position and a momentum. Any pair of 2-dimensional vectors, $\mathbf{a} = [a^\theta, a^p]^T$ and $\mathbf{b} = [b^\theta, b^p]^T$, can be used to describe the length and height of a parallelogram, whose oriented (i.e. positive) area is given by

$$\text{Area} = \left| \det \begin{pmatrix} a^p & b^p \\ a^\theta & b^\theta \end{pmatrix} \right| = |a^p b^\theta - a^\theta b^p| \quad (2.25)$$

If $d > 1$ then the total area, η , is defined as the sum of the areas induced by the pairs of position-momentum vectors, such that

$$\eta(\mathbf{a}, \mathbf{b}) \equiv \sum_{i=1}^d \left| \det \begin{pmatrix} a_i^p & b_i^p \\ a_i^\theta & b_i^\theta \end{pmatrix} \right| = \sum_{i=1}^d |a_i^p b_i^\theta - a_i^\theta b_i^p| \quad (2.26)$$

This can be written more succinctly in terms of the structure matrix \mathbf{J} (Equation 2.24)

$$\eta(\mathbf{a}, \mathbf{b}) \equiv \mathbf{a}^T \mathbf{J}^{-1} \mathbf{b} \quad (2.27)$$

This definition extends to classifying mappings as being symplectic. Let us consider a linear transformation, \mathbf{M} , of our vectors \mathbf{a} and \mathbf{b} ,

$$\eta(\mathbf{Ma}, \mathbf{Mb}) = (\mathbf{Ma})^T \mathbf{J}^{-1} (\mathbf{Mb}) \quad (2.28)$$

$$= \mathbf{a}^T \mathbf{M}^T \mathbf{J}^{-1} \mathbf{Mb} \quad (2.29)$$

If we want the areas enclosed by our original vectors and our newly transformed vectors to be equal, this implies the condition

$$\eta(\mathbf{Ma}, \mathbf{Mb}) = \eta(\mathbf{a}, \mathbf{b}) \quad (2.30)$$

and therefore

$$\mathbf{a}^T \mathbf{M}^T \mathbf{J}^{-1} (\mathbf{Mb}) = \mathbf{a}^T \mathbf{J}^{-1} \mathbf{b} \quad \Rightarrow \quad \mathbf{M}^T \mathbf{J}^{-1} \mathbf{M} = \mathbf{J}^{-1} \quad \forall \mathbf{a}, \mathbf{b} \quad (2.31)$$

This is the symplectic condition that we can use to prove the symplecticness of the flow map for a Hamiltonian system. In particular we want to check that the volume enclosed by points in nearby solutions is constant for each value of τ . The application of this symplectic condition extends to nonlinear mappings that can be approximated by a local linearisation with respect to the initial position and momentum variables (83). We must therefore show that the following condition holds for all τ ,

$$\left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right)^T \mathbf{J}^{-1} \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) = \mathbf{J}^{-1} \quad (2.32)$$

where $\mathbf{z} = [\boldsymbol{\theta}_0, \mathbf{p}_0]$ are the initial conditions for Hamilton's equations. This condition can easily be proven to be true. We firstly show that the derivative of the left hand side of Equation 2.32 is equal to 0, and therefore the left hand side must be equal to some constant. We then evaluate this constant by considering the particular case when $\tau = 0$, and conclude that this must therefore be the solution for all values of τ .

Writing Hamilton's equations (2.23) in terms of its flow map and differentiating with respect to \mathbf{z} we note that

$$\frac{d}{d\tau} \left[\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right] = \frac{\partial}{\partial \mathbf{z}} \left[\frac{d\Phi_\tau(\mathbf{z})}{d\tau} \right] \quad (2.33)$$

$$= \mathbf{J} H_{\mathbf{z}\mathbf{z}}(\Phi_\tau(\mathbf{z})) \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) \quad (2.34)$$

where $H_{\mathbf{z}\mathbf{z}}$ denotes the Hessian matrix of second partial derivatives of the Hamiltonian function H with respect to \mathbf{z} . We also note that the symplectic matrix \mathbf{J} has the following properties, $\mathbf{J}\mathbf{J} = 1$, $\mathbf{J}^{-1} = -\mathbf{J}$ and $\mathbf{J}^T = -\mathbf{J}$. Using these identities, we can show that the time derivative of the left hand side of Equation 2.32 is zero, which follows as

$$\begin{aligned} \frac{d}{d\tau} \left[\left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right)^T \mathbf{J}^{-1} \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) \right] &= \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right)^T \mathbf{J}^{-1} \left[\mathbf{J} H_{\mathbf{z}\mathbf{z}}(\Phi_\tau(\mathbf{z})) \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) \right] \\ &\quad + \left[\left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right)^T H_{\mathbf{z}\mathbf{z}}(\Phi_\tau(\mathbf{z})) \mathbf{J}^T \right] \mathbf{J}^{-1} \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) \\ &= \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right)^T H_{\mathbf{z}\mathbf{z}}(\Phi_\tau(\mathbf{z})) \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) \\ &\quad - \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right)^T H_{\mathbf{z}\mathbf{z}}(\Phi_\tau(\mathbf{z})) \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) \\ &= \mathbf{0} \end{aligned}$$

Finally, we make the observation that for $\tau = 0$, the flow map is simply an identity matrix and so condition 2.32 holds trivially, and therefore for all τ . This symplectic property is also equivalent to the Jacobian of the transition map having unit determinant. We can see this by considering the determinant of the symplectic condition

$$\det(\mathbf{J}^{-1}) = \det \left[\left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right)^T \mathbf{J}^{-1} \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) \right] \quad (2.35)$$

$$= \left[\det \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) \right]^2 \det(\mathbf{J}^{-1}) \quad (2.36)$$

Since $\det(\mathbf{J}^{-1}) \neq 0$ then the above implies that

$$\det \left(\frac{\partial \Phi_\tau(\mathbf{z})}{\partial \mathbf{z}} \right) = 1 \quad (2.37)$$

Given that the symplectic condition implies that the determinant of the Jacobian of a symplectic mapping is equal to one, we can consider the effect of this mapping on the volume, V , enclosed by a group of points by using the standard change of variables formula from multivariate calculus,

$$V(\Phi_\tau(U)) = \int_{\Phi_\tau(U)} dz_1 \dots dz_{2d} \quad (2.38)$$

$$= \int_U \left| \det \left(\frac{\partial \Phi_\tau}{\partial \mathbf{z}} \right) \right| dz_1 \dots dz_{2d} \quad (2.39)$$

$$= \int_U dz_1 \dots dz_{2d} \quad (2.40)$$

$$= V(U) \quad (2.41)$$

The volume enclosed by the group of transformed points is therefore equal to the volume enclosed by the original points and this identity is often referred to as Liouville's Theorem. This is equivalent to the divergence of the vector field described by Hamilton's equations being equal to zero (83), as the rate of change of the volume is zero. Therefore, assuming we can exactly calculate the Hamiltonian flow, we see

$$\nabla \cdot \left[\frac{d\boldsymbol{\theta}}{d\tau}, \frac{d\mathbf{p}}{d\tau} \right] = \frac{\partial}{\partial \boldsymbol{\theta}} \frac{d\boldsymbol{\theta}}{d\tau} + \frac{\partial}{\partial \mathbf{p}} \frac{d\mathbf{p}}{d\tau} \quad (2.42)$$

$$= \frac{\partial}{\partial \boldsymbol{\theta}} \frac{\partial H}{\partial \mathbf{p}} - \frac{\partial}{\partial \mathbf{p}} \frac{\partial H}{\partial \boldsymbol{\theta}} \quad (2.43)$$

$$= 0 \quad (2.44)$$

and so that the divergence is zero, as expected.

In the vast majority of cases however, analytic solutions of Hamilton’s equations are not available and one must resort to numerical integration methods based on a discretisation. Quite remarkably, it turns out that symplecticness can be preserved even for approximate discretised solutions to Hamilton’s equations, and this is vital for developing correct MCMC schemes. We will shortly see the impact of symplecticness on the overall numerical accuracy when integrating conservative Hamiltonian systems, however let us first look at the use of Hamiltonian systems for the purpose of simulating molecular dynamics in a physical context.

2.2 Molecular Dynamical Simulations

Molecular Dynamics aims at estimating the average properties of a molecular system as it evolves over time (3, 65). Such methods were originally developed for estimating equilibrium properties of materials and fluids and have the advantage over Monte Carlo methods that they also allow investigation of dynamic properties of the system, such as time dependent responses to perturbations (162), since the individual movements of molecules within the system are explicitly modelled in real time. This computational approach in many ways mirrors a real life experimental approach; a mathematical model is chosen to describe a complex system, its configuration is initialised at some sensible values, and the evolution of the ensemble is modelled explicitly by simulating the motion of each molecule according to the laws of classical mechanics, most conveniently in the form of Hamilton’s equations. As we have seen already, Hamilton’s equations are defined in continuous-time, however they do not generally admit analytic solutions and so we must resort to numerically solving them in a discrete form. The challenge is therefore to preserve at least the qualitative properties of the original continuous system.

Each molecule is described by a position vector and momentum vector. As this is a physical simulation, these vectors are both 3 dimensional and so each molecule is represented in a 6 dimensional phase space. The quantities of interest, which must be expressible as a function of position and momentum, are then computed as an average of the time evolution of the system states. As in real life it is often necessary to allow the system to reach an equilibrium state before accurate ergodic averages may be obtained.

The main computational cost of a Molecular Dynamics simulation lies in computing the force on each molecule, as opposed to the cost of integrating the equations themselves, since in principle each molecule exerts a force on every other one regardless of distance, and so the required calculations can become exponentially more expensive. This is particularly a problem when one considers that there may be thousands of molecules being modelled, although in practice there are approximations that can be made (65). The choice of integration scheme is also particularly important in Molecular Dynamics for a couple of reasons. Firstly there must be good energy preservation over the time period of the integration; the Hamiltonian is a conservative system and major fluctuations in the total energy can seriously affect the type of dynamics predicted, regardless of the accuracy of the model describing the forces involved. Later we will see that approximate energy conservation also plays an important role in setting the efficiency of a dynamical MCMC scheme. Secondly, it is desirable for the numerical scheme to be accurate for relatively large time steps, as this means that fewer force evaluations are required. We will see that in the context of MCMC, our force function will be replaced by a potentially expensive likelihood function, and so the same principles will apply.

The main concern regarding the use of molecular dynamics is the error associated with the integration of Hamilton's equations. Most of the commonly used integration schemes do not exactly preserve the total energy of the Hamiltonian, which raises the question of how accurate such simulations might be over longer periods of time. As an example, the stability of the solar system may be simulated by considering interactions of planets instead of molecules and using appropriate force calculations (190). In this case where the initial conditions are known precisely and the precise trajectory is of interest, as opposed to the average behaviour, energy preservation will be very important and small errors could seriously affect the results. In other applications however, where it is a much larger number of interacting objects that is of interest, the average behaviour may not be as strongly affected by small fluctuations in the total energy; in such cases we are not interested in calculating a precise trajectory since we will not know the precise initial positions and momentum of molecules, although we will at least have some idea of sensible starting values.

This estimation of average behaviour relies on the assumption that approximate simulation results using numerical integration schemes are generally close to the true

trajectories of the continuous time system during a particular time period of interest, despite the inevitable numerical integration errors. The use of symplectic integrators can help preserve the geometric qualities of the system, although it seems that the evidence for slow divergence from true trajectories is largely numerical and it is noted that this does not appear to have been explicitly proven, even for specific systems of interest (65). There is no doubt that Molecular Dynamics simulations are extremely useful for modelling the time evolution of many-body problems and are capable of reproducing the dynamical properties of many physical phenomena. Such simulations must be run with a careful monitoring of the total energy in the system and with an appropriate integration step size such that the average simulated dynamics may be trusted to represent the true average behaviour. We must bear in mind however that from a statistical point of view there is no mathematical guarantee of convergence to the true stationary distribution when simulating these discrete dynamics.

2.3 Methods of Numerical Integration

The idea of obtaining a numerical solution to a system of differential equations by employing a sequence of discrete calculations had already been considered in the 17th century; indeed Newton used such numerical methods to approximate solutions to his second law of motion, long before automated electronic computation became available. Numerical integration schemes can be derived by considering truncated Taylor expansions, which were formally introduced at the beginning of the 18th century, and it is perhaps quite remarkable that such schemes are still among the most popular and useful today. We shall briefly review some numerical methods for the integration of differential equations and focus in particular on those properties of that are important for obtaining solutions for conservative Hamiltonian systems. We will bear in mind our ultimate goal of employing them within the context of an MCMC scheme.

2.3.1 Euler's Method

We may consider the Taylor expansion of an infinitely differentiable function f at time τ with a timestep ϵ ,

$$f(\tau + \epsilon) \approx f(\tau) + \epsilon f'(\tau) + \frac{\epsilon^2}{2} f''(\tau) + O(\epsilon^3) \quad (2.45)$$

2.3 Methods of Numerical Integration

From this simple formula we can derive numerical schemes capable of solving the type of nonlinear differential equations we find expressed through Hamilton's equations. If we apply this Taylor expansion to a standard separable Hamiltonian system, expressed in the concatenated vector form $z = [\boldsymbol{\theta}, \mathbf{p}]$ (Equation 2.23), truncate all terms involving derivatives higher than first order and write out in terms of position and momentum variables, we obtain Euler's method.

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \epsilon \frac{d\boldsymbol{\theta}(\tau)}{d\tau} \quad (2.46)$$

$$= \boldsymbol{\theta}(\tau) + \epsilon \mathbf{M}^{-1} \mathbf{p}(\tau) \quad (2.47)$$

$$\mathbf{p}(\tau + \epsilon) = \mathbf{p}(\tau) + \epsilon \frac{d\mathbf{p}(\tau)}{d\tau} \quad (2.48)$$

$$= \mathbf{p}(\tau) - \epsilon \frac{\partial E_p(\boldsymbol{\theta}(\tau))}{\partial \boldsymbol{\theta}} \quad (2.49)$$

We can write this in terms of sequentially updating the n th step of our integration scheme as,

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \epsilon \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^n, \mathbf{p}^n) \quad (2.50)$$

$$\mathbf{p}^{n+1} = \mathbf{p}^n - \epsilon \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^n, \mathbf{p}^n) \quad (2.51)$$

where we use the short-hand convention of using $\nabla_{\mathbf{p}} H$ to denote the partial derivative of the Hamiltonian with respect to the momentum; in this section we write out all other derivatives in full for clarity. This method is said to be first order accurate, which is clear from its construction. More generally, the magnitude of any errors associated with an integration scheme can be determined by comparing the timestep updates with the Taylor expansion around the current point. Unfortunately this Euler method produces very poor results even for simple Hamiltonian systems, with both the dynamics and the total energy rapidly diverging from their true trajectories, as demonstrated in Figure 2.1.

2.3.2 Method of Splitting

When considering the integration of separable Hamiltonians we can use the idea of *splitting* to obtain numerical schemes that are better behaved, in particular with respect

2.3 Methods of Numerical Integration

to energy conservation. A separable Hamiltonian may be represented as a sum of two independent Hamiltonians; one a function of momentum and the other a function of position. The original Hamiltonian may then be numerically integrated by considering the composition of two separate flow maps, one evolving with respect to momentum and the other with respect to position. Let us consider the Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^2 + E_p(\boldsymbol{\theta})$, which we may write as a sum of two independent Hamiltonians $H(\boldsymbol{\theta}, \mathbf{p}) = H_1(\mathbf{p}) + H_2(\boldsymbol{\theta})$, where $H_1 = \frac{1}{2}\mathbf{p}^2$ and $H_2 = E_p(\boldsymbol{\theta})$. We can see that Hamilton's equations for H_1 are

$$\frac{d\boldsymbol{\theta}}{d\tau} = \mathbf{0}, \quad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial E_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (2.52)$$

which induce the flow map

$$\Phi_{\epsilon, H_1} \begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{p} \end{pmatrix} = \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{p} - \epsilon \frac{\partial E_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{bmatrix} \quad (2.53)$$

Similarly for H_2 , Hamilton's equations are simply

$$\frac{d\boldsymbol{\theta}}{d\tau} = \mathbf{p}, \quad \frac{d\mathbf{p}}{d\tau} = \mathbf{0} \quad (2.54)$$

which induce the flow map

$$\Phi_{\epsilon, H_2} \begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{p} \end{pmatrix} = \begin{bmatrix} \boldsymbol{\theta} + \epsilon \mathbf{p} \\ \mathbf{p} \end{bmatrix} \quad (2.55)$$

Taking the composition of these two flow maps, we obtain the symplectic Euler method of solving Hamilton's equations, which we denote Φ_A ,

$$\Phi_A = \Phi_{\epsilon, H_1} \circ \Phi_{\epsilon, H_2} \begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{p} \end{pmatrix} = \Phi_{\epsilon, H_1} \begin{bmatrix} \boldsymbol{\theta} + \epsilon \mathbf{p} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta} + \epsilon \mathbf{p} \\ \mathbf{p} - \epsilon \frac{\partial E_p(\boldsymbol{\theta} + \epsilon \mathbf{p})}{\partial \boldsymbol{\theta}} \end{bmatrix} \quad (2.56)$$

Writing this in terms of the original Hamiltonian we may update the position and momentum vectors as follows,

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \epsilon \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^n, \mathbf{p}^n) \quad (2.57)$$

$$\mathbf{p}^{n+1} = \mathbf{p}^n - \epsilon \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{n+1}, \mathbf{p}^n) \quad (2.58)$$

Reversing the composition of the flow maps gives an alternative symplectic Euler method, Φ_B , whereby the momentum values are calculated first and employed in the update of the position vector,

$$\Phi_B = \Phi_{\epsilon, H_2} \circ \Phi_{\epsilon, H_1} \left(\begin{array}{c} \boldsymbol{\theta} \\ \mathbf{p} \end{array} \right) = \Phi_{\epsilon, H_2} \left[\begin{array}{c} \boldsymbol{\theta} \\ \mathbf{p} - \epsilon \frac{\partial E_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{array} \right] = \left[\begin{array}{c} \boldsymbol{\theta} + \epsilon \left(\mathbf{p} - \epsilon \frac{\partial E_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \\ \mathbf{p} - \epsilon \frac{\partial E_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{array} \right]$$

Again we can write this in terms of sequentially updating the position and momentum vectors as follows,

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \epsilon \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^n, \mathbf{p}^{n+1}) \quad (2.59)$$

$$\mathbf{p}^{n+1} = \mathbf{p}^n - \epsilon \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^n, \mathbf{p}^n) \quad (2.60)$$

Both of these algorithms are first order methods and turn out to be superior to the Euler scheme applied directly to the joint vector z . An illustrative comparison is shown in Figure 2.1. This approach of splitting Hamiltonians easily allows higher order schemes to be constructed; we can simply by split the Hamiltonian into more components with fractional time steps, such that the overall time step of their composition is equal to 1, for both potential and kinetic energy terms (122).

An important example of this is the splitting $H = H_1 + H_2 + H_3$, where $H_1 = \frac{1}{2}E_p = \frac{1}{2}\mathbf{p}^2$, $H_2 = E_k$ and $H_3 = \frac{1}{2}E_p = \frac{1}{2}\mathbf{p}^2$, which corresponds to the composition of the following flow maps,

$$\Phi_{\frac{\epsilon}{2}, H_1} \circ \Phi_{\epsilon, H_2} \circ \Phi_{\frac{\epsilon}{2}, H_3} \quad (2.61)$$

This scheme may be used for separable Hamiltonians, since we assumed that the potential energy was only a function of the position, $\boldsymbol{\theta}$, and that the kinetic energy was only a function of the momentum, \mathbf{p} . Writing this composition out in full, we obtain the Leapfrog algorithm,

2.3 Methods of Numerical Integration

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^n - \frac{\epsilon}{2} \frac{\partial E_p(\boldsymbol{\theta}^n)}{\partial \boldsymbol{\theta}} \quad (2.62)$$

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \epsilon \mathbf{p}^{n+\frac{1}{2}} \quad (2.63)$$

$$\mathbf{p}^{n+1} = \mathbf{p}^{n+\frac{1}{2}} - \frac{\epsilon}{2} \frac{\partial E_p(\boldsymbol{\theta}^{n+1})}{\partial \boldsymbol{\theta}} \quad (2.64)$$

This integration scheme is a second order method and is commonly used for integrating conservative systems defined in terms of a separable Hamiltonian, due to its low divergence from true trajectories and relatively good energy preservation over time. This Leapfrog scheme may also be derived from the two symplectic Euler schemes, by considering the composition of the first with the second scheme, i.e. a half step using Euler-B, then a half step using Euler-A,

$$\Phi_{\frac{\epsilon}{2}, H_A} \circ \Phi_{\frac{\epsilon}{2}, H_B} \quad (2.65)$$

which results in the following updating scheme

$$\boldsymbol{\theta}^{n+\frac{1}{2}} = \boldsymbol{\theta}^n + \frac{\epsilon}{2} \mathbf{p}^{n+\frac{1}{2}} \quad (2.66)$$

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^n - \frac{\epsilon}{2} \frac{\partial E_p(\boldsymbol{\theta}^n)}{\partial \boldsymbol{\theta}} \quad (2.67)$$

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^{n+\frac{1}{2}} + \frac{\epsilon}{2} \mathbf{p}^{n+\frac{1}{2}} \quad (2.68)$$

$$\mathbf{p}^{n+1} = \mathbf{p}^{n+\frac{1}{2}} - \frac{\epsilon}{2} \frac{\partial E_p(\boldsymbol{\theta}^{n+1})}{\partial \boldsymbol{\theta}} \quad (2.69)$$

This may be simplified by substituting the first equation for $\boldsymbol{\theta}^{n+\frac{1}{2}}$ into the third equation. We may obtain a more general version of the Leapfrog algorithm by rewriting it in terms of partial derivatives of the original Hamiltonian, such that

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^n - \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^n, \mathbf{p}^{n+\frac{1}{2}}) \quad (2.70)$$

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \frac{\epsilon}{2} \left[\nabla_{\mathbf{p}} H(\boldsymbol{\theta}^n, \mathbf{p}^{n+\frac{1}{2}}) + \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^{n+1}, \mathbf{p}^{n+\frac{1}{2}}) \right] \quad (2.71)$$

$$\mathbf{p}^{n+1} = \mathbf{p}^{n+\frac{1}{2}} - \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{n+1}, \mathbf{p}^{n+\frac{1}{2}}) \quad (2.72)$$

This is the Störmer-Verlet or Generalised Leapfrog integration scheme and its equations are *explicit* for separable Hamiltonians. For non-separable Hamiltonians, however, the first two equations are defined *implicitly*; the same unknown term appears on both sides of the equation. This must then be solved with the aid of an additional numerical scheme, such as fixed point iterations. We shall discuss this further in Chapter 3 and for now consider only separable Hamiltonian systems.

2.3.3 Illustrative Example: Parameters of a Gaussian Distribution

We can see the performance of such numerical integration schemes by considering a Hamiltonian system in which the potential energy is given by the negative log-likelihood of a simple statistical model, in this case a Gaussian distribution. We shall present more detail about this statistical interpretation of a Hamiltonian later in Section 2.4, but for the time being let us simply consider $N = 30$ observations drawn from a Gaussian distribution $\mathcal{N}(\mathbf{y}|\mu = 0, \sigma = 10)$, such that our parameter (position) space is 2-dimensional. The log-likelihood for such a model follows as

$$L(\mathbf{y}|\mu, \sigma) \propto -N \log(\sigma) - \frac{1}{2\sigma^2} \sum_n^N (y_n - \mu)^2 \quad (2.73)$$

In order to solve Hamilton's equations, we also need the derivatives of this expression with respect to each of the parameters μ and σ . These follow as

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2} \sum_n^N (y_n - \mu) \quad (2.74)$$

$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_n^N (y_n - \mu)^2 \quad (2.75)$$

$$(2.76)$$

We initialise our parameters at $\mu = 2$ and $\sigma = 10$, and initialise the momentum variables at $\mathbf{p} = [1, 1]$. We then integrate the appropriate Hamiltonian for 50 iterations using the Euler, symplectic Euler and Leapfrog schemes with a stepsize of 0.1. Figure 2.1 shows both the trajectories in the parameter space and the fluctuations in the total energy H . The Euler scheme has very poor accuracy overall. The symplectic

Euler scheme has much better numerical accuracy for the position variables, however the total energy fluctuates rather dramatically. In contrast, the higher order Leapfrog scheme has visibly much better accuracy, both in terms of the position variables and preservation of total energy.

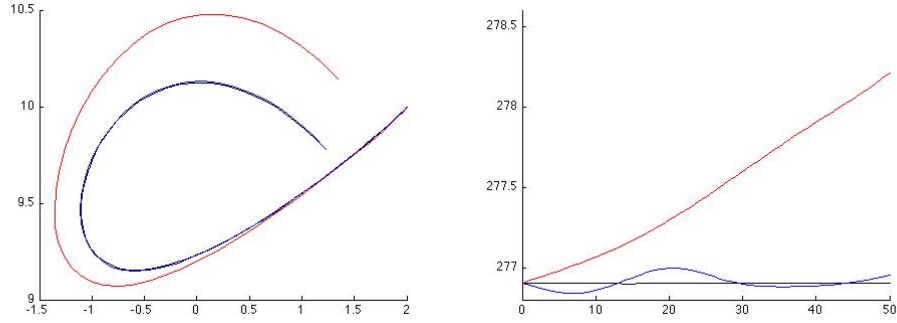


Figure 2.1: Comparison of integration schemes for Hamilton's Equations - The left hand plot shows the trajectories of the position variables using a Hamiltonian whose potential energy is defined by the negative log-likelihood of a simple Gaussian statistical model. The right hand side shows the total energy of the Hamiltonian at each step of the integration scheme. In both plots the Euler scheme output is red, the symplectic Euler scheme output is blue and the Leapfrog scheme is black.

2.3.4 Properties of Integration Schemes

It is now interesting to investigate the fundamental properties these integration schemes possess that make them particularly well suited for solving the Hamiltonian dynamics of a conservative system. The Leapfrog integrator is another example of a *symplectic* map, albeit a discrete map in contrast to the continuous symplectic map induced by Hamilton's equations, and this turns out to be the key property that enables accurate results to be calculated for such systems. As we saw earlier in this chapter, we can characterise symplecticity by considering the flow map of a transformation. An integration scheme with flow map Φ from \mathbb{R}^{2d} to \mathbb{R}^{2d} is volume preserving if the symplectic condition holds

$$\left[\frac{\partial}{\partial \mathbf{z}} \Phi(\mathbf{z}) \right]^T \mathbf{J}^{-1} \left[\frac{\partial}{\partial \mathbf{z}} \Phi(\mathbf{z}) \right] = \mathbf{J}^{-1} \quad (2.77)$$

2.3 Methods of Numerical Integration

where \mathbf{J} is the symplectic structure matrix, which is given by representing the Hamiltonian in its canonical form (Equation 2.23). We can now show that the standard Euler scheme is not symplectic, which explains its poor accuracy when integrating a symplectic Hamiltonian system. We may calculate the derivative of the flow map of the Euler method,

$$\frac{\partial}{\partial \mathbf{z}} \Phi(\mathbf{z}) = \begin{bmatrix} \frac{\partial \theta^{n+1}}{\partial \theta^n} & \frac{\partial \theta^{n+1}}{\partial \mathbf{p}^n} \\ \frac{\partial \mathbf{p}^{n+1}}{\partial \theta^n} & \frac{\partial \mathbf{p}^{n+1}}{\partial \mathbf{p}^n} \end{bmatrix} = \begin{bmatrix} \mathbf{I} + \epsilon H_{\theta \mathbf{p}}(\theta^n, \mathbf{p}^n) & \epsilon H_{\mathbf{p} \mathbf{p}}(\theta^n, \mathbf{p}^n) \\ -\epsilon H_{\theta \theta}(\theta^n, \mathbf{p}^n) & \mathbf{I} - \epsilon H_{\mathbf{p} \theta}(\theta^n, \mathbf{p}^n) \end{bmatrix}$$

Noting that $\mathbf{J}^{-1} = -\mathbf{J}$, and denoting the partial derivative of $H(\theta^n, \mathbf{p}^n)$ with respect to θ by H_{θ} , we see that

$$\begin{aligned} \left[\frac{\partial}{\partial \mathbf{z}} \Phi(\mathbf{z}) \right]^T \mathbf{J}^{-1} \left[\frac{\partial}{\partial \mathbf{z}} \Phi(\mathbf{z}) \right] &= \begin{bmatrix} \mathbf{I} + \epsilon H_{\theta \mathbf{p}} & -\epsilon H_{\theta \theta} \\ \epsilon H_{\mathbf{p} \mathbf{p}} & \mathbf{I} - \epsilon H_{\mathbf{p} \theta} \end{bmatrix} \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} + \epsilon H_{\theta \mathbf{p}} & \epsilon H_{\mathbf{p} \mathbf{p}} \\ -\epsilon H_{\theta \theta} & \mathbf{I} - \epsilon H_{\mathbf{p} \theta} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{I} + 2\epsilon H_{\theta \mathbf{p}} \\ \mathbf{I} - \epsilon^2 H_{\theta \mathbf{p}} H_{\mathbf{p} \theta} + \epsilon^2 H_{\mathbf{p} \mathbf{p}} H_{\theta \theta} & \mathbf{0} \end{bmatrix} \\ &\neq \mathbf{J}^{-1} \end{aligned}$$

and so the standard Euler scheme is not symplectic. It can be verified that the symplectic Euler schemes are indeed, as their name suggests, symplectic by implicitly differentiating them and substituting the result into Equation 2.77 to show that the symplectic condition holds. The power of the splitting method to produce new schemes lies in the fact that the composition of symplectic flow maps preserves the symplectic property. This follows straightforwardly by considering two symplectic maps Φ_1 and Φ_2 , and noting that their composition also satisfies the symplectic condition since, using the chain rule,

$$\left[\frac{\partial \Phi_1}{\partial \mathbf{z}} \frac{\partial \Phi_2}{\partial \mathbf{z}} \right]^T \mathbf{J}^{-1} \left[\frac{\partial \Phi_1}{\partial \mathbf{z}} \frac{\partial \Phi_2}{\partial \mathbf{z}} \right] = \left[\frac{\partial \Phi_2}{\partial \mathbf{z}} \right]^T \left[\frac{\partial \Phi_1}{\partial \mathbf{z}} \right]^T \mathbf{J}^{-1} \left[\frac{\partial \Phi_1}{\partial \mathbf{z}} \right] \left[\frac{\partial \Phi_2}{\partial \mathbf{z}} \right] \quad (2.78)$$

$$= \left[\frac{\partial \Phi_2}{\partial \mathbf{z}} \right]^T \mathbf{J}^{-1} \left[\frac{\partial \Phi_2}{\partial \mathbf{z}} \right] \quad (2.79)$$

$$= \mathbf{J}^{-1} \quad (2.80)$$

From this it follows that the Generalised Leapfrog scheme must also be symplectic, as it may be described as a composition of the two symplectic Euler schemes. Another

observation we may make is that Newton's equations are time-reversible, therefore we may wish to have an integration scheme that is also time-reversible. Time-reversibility also turns out to be very useful for defining a valid MCMC scheme as it allows the detailed balance condition to be easily satisfied. If we reverse the timestep for the Leapfrog method, we see that we end up back at exactly the same starting point, and so this method is time-reversible. Let us consider one iteration of the Generalised Leapfrog method as a straightforward example. We perform one iteration of this integrator with a time step ϵ , moving from a point $(\boldsymbol{\theta}, \mathbf{p})$ to $(\boldsymbol{\theta}^*, \mathbf{p}^*)$, via a half momentum step denoted by $\mathbf{p}^{\frac{1}{2}}$. The equations follow as

$$\mathbf{p}^{\frac{1}{2}} = \mathbf{p} - \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}^{\frac{1}{2}}) \quad (2.81)$$

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} + \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}^{\frac{1}{2}}) + \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^*, \mathbf{p}^{\frac{1}{2}}) \quad (2.82)$$

$$\mathbf{p}^* = \mathbf{p}^{\frac{1}{2}} - \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^*, \mathbf{p}^{\frac{1}{2}}) \quad (2.83)$$

Let us now consider the backwards trajectory by once again applying one iteration of the Generalised Leapfrog method, negating the timestep and starting at $(\boldsymbol{\theta}^*, \mathbf{p}^*)$,

$$\hat{\mathbf{p}}^{\frac{1}{2}} = \mathbf{p}^* + \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^*, \hat{\mathbf{p}}^{\frac{1}{2}}) \quad (2.84)$$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* - \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^*, \hat{\mathbf{p}}^{\frac{1}{2}}) - \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\hat{\boldsymbol{\theta}}, \hat{\mathbf{p}}^{\frac{1}{2}}) \quad (2.85)$$

$$\hat{\mathbf{p}} = \hat{\mathbf{p}}^{\frac{1}{2}} + \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\hat{\boldsymbol{\theta}}, \hat{\mathbf{p}}^{\frac{1}{2}}) \quad (2.86)$$

where the newly obtained position and momentum values are denoted by $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{p}}$ respectively. We can see that by rearranging Equation 2.84 (the first step of the integrator with the time reversed) we can compare with Equation 2.83 (the last step of the integrator in forward time). Since the only unknown variable is $\hat{\mathbf{p}}^{\frac{1}{2}}$, we conclude that $\hat{\mathbf{p}}^{\frac{1}{2}} = \mathbf{p}^{\frac{1}{2}}$. We can then rearrange Equation 2.85 and see that we can directly compare with Equation 2.82. Once again, since the only unknown variable is $\hat{\boldsymbol{\theta}}$, we conclude that $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. Finally, we can rearrange Equation 2.86 and, comparing with Equation 2.81, we conclude that $\hat{\mathbf{p}} = \mathbf{p}$, since $\hat{\mathbf{p}}$ is the only unknown variable. We note that we can also instead negate \mathbf{p} and integrate forwards in time to obtain the same results, due to the symmetry of the Hamiltonian. By solving the implicit equations of the Generalised

Leapfrog method, we can obtain **exactly** the same reverse path by negating t , and so our method is reversible.

Molecular Dynamics simulations are often run by integrating a Hamiltonian system using an accurate symplectic numerical scheme, such as the Leapfrog method we have just seen. This particular method is popular because of its simple form and relatively stable numerical properties. We may obtain even more accurate results by considering higher order symplectic methods and there exists a large literature examining geometric integrators for a wide variety of Hamiltonian systems in many different contexts (83, 122). Much research has also focused on investigating their numerical properties, particularly with respect to stability and the rate of divergence from true solutions (122).

Regardless of the numerical method used however, such approaches are not guaranteed to result in samples from the true stationary distribution, and often it is difficult to quantify by exactly how much the results deviate from the true solution. This problem may be solved by considering a statistical perspective and combining Molecular Dynamics approaches with a Metropolis-Hastings acceptance step to correct for discretisation errors.

2.4 Hamiltonian Monte Carlo

We can consider Hamiltonian Monte Carlo from two perspectives; from a Molecular Dynamics viewpoint we see HMC as a method of correcting the errors introduced by a discretisation of Hamilton's equations. From a statistics point of view, we see the use of Hamiltonian dynamics as an effective proposal mechanism for MCMC, such that the proposal states will be far from the current state and accepted with high probability.

Molecular dynamics was put on statistically more solid ground by Duane et al. (57), who combined its ideas with related concepts from MCMC methods. Until this point these two approaches had been considered only separately, even though there is a great overlap in the type of problem they are trying to solve. The approach is to employ a standard discretised molecular dynamics scheme and correct for the integration error at the end of the simulation step by accepting or rejecting the move according to a Metropolis-Hastings ratio involving the total energy. The main advantage of this approach is that we need no longer be concerned with obtaining great accuracy in

the integration step, as the samples drawn from this statistically corrected version of molecular dynamics will still be guaranteed to be from the desired target distribution. The connection with more general statistical models was then made by Neal (145), who demonstrated that the Hamiltonian could be defined in terms of the parameters of a statistical model with the potential energy given by its negative log-likelihood.

Let us now focus on Hamiltonian dynamics from this statistical point of view. We may interpret the coordinate vector $\boldsymbol{\theta}$ as a random variable $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$, and we interpret the momentum $\mathbf{p} \in \mathbb{R}^D$ as an independent auxiliary variable with density $p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$. The joint density follows in factorised form as $p(\boldsymbol{\theta}, \mathbf{p}) = p(\boldsymbol{\theta})p(\mathbf{p}) = p(\boldsymbol{\theta})\mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$, and the Hamiltonian has a natural interpretation as the negative logarithm of this joint likelihood

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2} \log((2\pi)^D |\mathbf{M}|) + \frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} \quad (2.87)$$

where we have added an additional normalisation term for the Gaussian momentum, such that the marginals correspond to the desired target distributions. We can see this by integrating the negative exponential of the Hamiltonian with respect to \mathbf{p}

$$p(\boldsymbol{\theta}) \propto \int \exp(-H(\boldsymbol{\theta}, \mathbf{p})) d\mathbf{p} = \frac{\exp\{\mathcal{L}(\boldsymbol{\theta})\}}{\sqrt{(2\pi)^D |\mathbf{M}|}} \int \exp\left\{-\frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}\right\} d\mathbf{p} \quad (2.88)$$

$$= \exp\{\mathcal{L}(\boldsymbol{\theta})\} \quad (2.89)$$

and by integrating the Hamiltonian with respect to $\boldsymbol{\theta}$ we obtain

$$p(\mathbf{p}) \propto \int \exp(-H(\boldsymbol{\theta}, \mathbf{p})) d\boldsymbol{\theta} = \frac{\exp\left\{-\frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}\right\}}{\sqrt{(2\pi)^D |\mathbf{M}|}} \int \exp\{\mathcal{L}(\boldsymbol{\theta})\} d\boldsymbol{\theta} \quad (2.90)$$

$$= \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M}) \quad (2.91)$$

which is the density of the momentum $p(\mathbf{p}|\mathbf{0}, \mathbf{M})$. The mass matrix in the Hamiltonian therefore corresponds to the covariance matrix of our auxiliary variable. As before, the time evolution of this system may be described exactly by Hamilton's equations and we note that these simply involve the score functions with respect to $\boldsymbol{\theta}$ and \mathbf{p} of the joint density $H(\boldsymbol{\theta}, \mathbf{p})$ as follows

$$\frac{d\boldsymbol{\theta}}{d\tau} = \nabla_{\mathbf{p}}H(\boldsymbol{\theta}, \mathbf{p}) = \mathbf{M}^{-1}\mathbf{p} \quad (2.92)$$

$$\frac{d\mathbf{p}}{d\tau} = -\nabla_{\boldsymbol{\theta}}H(\boldsymbol{\theta}, \mathbf{p}) = \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) \quad (2.93)$$

where we denote the score function using the more succinct nabla notation, which we now use for the rest of this chapter. We may therefore obtain samples from our joint distribution $p(\boldsymbol{\theta}, \mathbf{p})$ by simulating points from this system according to Hamiltonian dynamics. Except in elementary cases, there are generally no analytic solutions to these equations of motion, and so for practical applications we must resort to numerical methods to obtain approximate solutions, as we saw in the previous section. Any discretisation introduces some integration error, and consequently the samples we obtain are no longer from the target distribution we are interested in. This is addressed by embedding the HMC proposals within a Metropolis-Hastings algorithm.

We may integrate Hamilton's equations using the Leapfrog integrator introduced in the previous section. This integration method relies on the fact that the Hamiltonian is separable and in statistical terms this means that the joint distribution must be factorisable. The Leapfrog integrator is as follows

$$\mathbf{p}(\tau + \epsilon/2) = \mathbf{p}(\tau) + (\epsilon/2)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\tau)) \quad (2.94)$$

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \epsilon\mathbf{M}^{-1}\mathbf{p}(\tau + \epsilon/2) \quad (2.95)$$

$$\mathbf{p}(\tau + \epsilon) = \mathbf{p}(\tau + \epsilon/2) + (\epsilon/2)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}(\tau + \epsilon)) \quad (2.96)$$

Due to the volume preserving property of the integrator, the determinant of the Jacobian matrix for the mapping defined by Φ_{τ} need not be taken into account in the Hastings ratio for calculating the acceptance probability. Therefore for a mapping $(\boldsymbol{\theta}, \mathbf{p}) \mapsto (\boldsymbol{\theta}^*, \mathbf{p}^*)$ obtained from a number of Leapfrog integration steps, the corresponding acceptance probability is $\min[1, \exp\{-H(\boldsymbol{\theta}^*, \mathbf{p}^*) + H(\boldsymbol{\theta}, \mathbf{p})\}]$ and due to the reversibility of the dynamics, the joint density and hence the marginals, $p(\boldsymbol{\theta})$ and $p(\mathbf{p})$, are invariant (148). We can explicitly see how detailed balance is satisfied,

$$p(\mathbf{x}_i)T(\mathbf{x}_j|\mathbf{x}_i) = p(\mathbf{x}_j)T(\mathbf{x}_i|\mathbf{x}_j), \forall i, j \quad (2.97)$$

Considering a small volume of phase space $\mathbf{x}_i = [\boldsymbol{\theta}_i, \mathbf{p}_i]$, and all other small volumes of phase space $\mathbf{x}_j = [\boldsymbol{\theta}_j, \mathbf{p}_j]$ to which \mathbf{x}_i can be mapped via a sequence of simulated Hamiltonian dynamics steps using a symplectic, time-reversible integrator. Firstly, since the transformation is volume preserving, the volume at \mathbf{x}_i is the same as the volume at \mathbf{x}_j . Secondly, since the mapping is time-reversible the probability of moving from \mathbf{x}_i to \mathbf{x}_j is the same as the reverse move. We then see that Equation 2.97 is satisfied for the joint density given by the Hamiltonian since,

$$p(\mathbf{x}_i)T(\mathbf{x}_j|\mathbf{x}_i) = \frac{\exp(-H(\mathbf{x}_i))}{z} \min[1, \exp(-H(\mathbf{x}_j)) - \exp(-H(\mathbf{x}_i))] \quad (2.98)$$

$$= \frac{\exp(-H(\mathbf{x}_j))}{z} \min[1, \exp(-H(\mathbf{x}_i)) - \exp(-H(\mathbf{x}_j))] \quad (2.99)$$

$$= p(\mathbf{x}_j)T(\mathbf{x}_i|\mathbf{x}_j) \quad (2.100)$$

where z is the unknown normalising constant. From a practical point of view, the acceptance probability of such proposal steps can be directly controlled by limiting the integration error through the choice of an appropriately small step size, ϵ . In practice this can either be fixed or drawn from some distribution; the latter might be helpful when the Markov chain starts far from the mode in the tails of the stationary distribution, where a different stepsize may be needed to get accurate trajectories (148).

For separable Hamiltonians, the Leapfrog integrator therefore provides a deterministic proposal mechanism; given the current parameters of our statistical model, $\boldsymbol{\theta}$, and random momentum values, \mathbf{p} , that are drawn exactly from the marginal distribution describing the kinetic energy, we obtain proposed values $\boldsymbol{\theta}^*$ and \mathbf{p}^* via numerical integration. These proposed values are accepted or rejected to ensure convergence to the correct stationary distribution, and so this Hamiltonian Monte Carlo method produces a time-reversible Markov chain that is ergodic and satisfies detailed balance (148). We can see that the stationary marginal distribution is indeed our target distribution of interest $\mathbf{p}(\boldsymbol{\theta})$, which may be obtained by simply disregarding the momentum samples, since $p(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}, \mathbf{p})$.

Hamiltonian Monte Carlo promises to offer more efficient sampling from high dimensional probability distributions by effectively reducing the amount of random walk present in the parameter values being proposed and exploiting the first order geometric information of the target density with respect to the model parameters. This has indeed

been shown to be the case for relatively simple, albeit high-dimensional, multivariate normal distributions, however there has been relatively little application to more complex statistical models, with the notable exception of Neal (146). We believe the reason for this lies in the amount of tuning that is often required to obtain reasonable mixing and rates of acceptance, although there do exist heuristics for certain classes of models used for linear and nonlinear regression (146). The two main parameters that require tuning are the number of leapfrog steps, N , and the size of each leapfrog step, ϵ . Setting different leapfrog stepsizes along different directions can be equivalently encoded in the mass matrix M (146, 148), which we will look at in more detail in Chapter 3. The use of exploratory runs of an MCMC sampler has been suggested (84) to obtain initial estimates of the target distribution and inform step sizes, however there is the obvious associated computational cost and the fact that this may not be feasible for very complex, high dimensional and potentially multimodal distributions.

2.5 Metropolis-adjusted Langevin Algorithm

An alternative approach to simulating Molecular Dynamics is to use a Langevin diffusion which describes the random movement of molecules in terms of a Brownian motion. Let us consider a random variable $\boldsymbol{\theta} \in \mathbb{R}^D$ with probability density $p(\boldsymbol{\theta})$ and denote the log density as $\mathcal{L}(\boldsymbol{\theta}) \equiv \log p(\boldsymbol{\theta})$. A Langevin diffusion is defined by the continuous-time stochastic differential equation (SDE)

$$d\boldsymbol{\theta}(t) = \frac{1}{2} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(t)) dt + d\mathbf{b}(t) \quad (2.101)$$

where \mathbf{b} denotes a D -dimensional Brownian motion. It is assumed that $p(\boldsymbol{\theta})$ is everywhere non-zero and differentiable, such that $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(t))$ is suitably well defined.

It has been shown that this continuous-time SDE converges to a unique invariant stationary distribution (105), however the discretised form of the SDE, the Unadjusted Langevin Diffusion (151), may not always converge and indeed may end up converging to a completely different stationary distribution. Despite this potential problem, the Langevin diffusion was employed in molecular dynamics in the late 1970s (172, 199) as a means of simulating molecular interactions, with further applications to areas of physics

2.5 Metropolis-adjusted Langevin Algorithm

appearing throughout the next decade (15, 42). A first order Euler discretisation of the SDE gives the following proposal mechanism

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \frac{\epsilon^2}{2} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^n) + \epsilon \mathbf{z}^n \quad (2.102)$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$ and ϵ is the integration stepsize. With this naive approximation, convergence to the invariant distribution $p(\boldsymbol{\theta})$ is no longer guaranteed due to the first-order integration error introduced by the finite step size ϵ . In a similar manner to Duane et al. (57), a simple modification of the Unadjusted Langevin Diffusion was proposed to circumvent this problem; we may use the discretised diffusion as a proposal step within a Metropolis-Hastings algorithm (80, 104) in order to correct any numerical errors, just as we did for a discretised Hamiltonian. This Metropolis Adjusted Langevin Algorithm (MALA) is now guaranteed to converge to the correct stationary distribution. Let us denote

$$\boldsymbol{\mu}(\boldsymbol{\theta}^n, \epsilon) = \boldsymbol{\theta}^n + \frac{\epsilon^2}{2} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^n) \quad (2.103)$$

then the discrete form of the SDE (Equation 2.101) defines a proposal density

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\mu}(\boldsymbol{\theta}^n, \epsilon), \epsilon^2 \mathbf{I}) \quad (2.104)$$

with acceptance probability of the standard form, $\min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^n|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^n)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n)\}$.

This “Metropolised” version of a discrete Langevin diffusion has been theoretically examined in (169, 171), where the authors investigate the rate of convergence under different conditions on the tails of the target distribution. The optimal scaling ϵ for MALA has also been theoretically analysed in the limit as the dimension $D \rightarrow \infty$ for factorisable $p(\boldsymbol{\theta})$ (168) and the optimal value of the acceptance rate has been calculated to be around 0.574, although in practice there may be reasonable performance using an acceptance rate of between 40% and 70%. These asymptotic results only hold once the Markov chain has reached stationarity, and more recently there has been investigation into the optimal scaling for the transient regimes, in which the chain has not yet reached equilibrium (38).

2.5 Metropolis-adjusted Langevin Algorithm

As the noise term \mathbf{z} is an isotropic standardised Normal variate, MALA is implicitly defined in a Euclidean geometry with a basis over this space defined canonically in terms of the parameters of the model. Proposed movement in each of these directions is equally likely, however we note that the probability density may change at dramatically different rates in each of these directions, so although the drift term in the proposal mechanism for MALA in Equation 2.104 is defined in terms of a Euclidean form of the gradient information, it is clear that an isotropic diffusion will be inefficient for strongly correlated variables $\boldsymbol{\theta}$ with widely differing variances, since the step size is effectively forced to accommodate the variate with the smallest variance. This issue can be partially addressed (38) by employing a pre-conditioning matrix \mathbf{M} such that

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \frac{\epsilon^2}{2} \mathbf{M}^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^n) + \epsilon \mathbf{M}^{-1} \mathbf{z}^n \quad (2.105)$$

The use of a pre-conditioning matrix is simply a linear transformation of the parameters, which may be interpreted as a change of basis. It might be hoped that this change of basis results in the transformed space being less correlated and easier to explore, allowing for larger proposal steps to be accepted. The tuning of MALA via the careful choice of step size and pre-conditioning matrix often plays a pivotal role in obtaining an efficient Markov chain with low correlation. Although it has proved useful to use pre-conditioning based on some estimate of the second order structure of the target density (116), it is often unclear how the pre-conditioning matrix should be defined in any principled manner. In particular we note that a particular choice of pre-conditioning may well be inappropriate for different regions of parameter space, and in addition different scalings may be necessary for the transient and stationary regimes of the Markov process as demonstrated in (38). We investigate this further in Chapter 4 with the example of a log-Gaussian Cox model.

There is an interesting connection between HMC and MALA; if we consider a single Leapfrog step, we see that by combining the update equations we obtain a single equation of the form,

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \frac{\epsilon^2}{2} \mathbf{M}^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(\tau)) + \epsilon \mathbf{M}^{-1} \mathbf{p}(\tau) \quad (2.106)$$

which can be interpreted as a discretised pre-conditioned Langevin diffusion as employed in MALA. We can now see that HMC will have similar issues to MALA with respect to choosing a suitable basis. In the case of HMC, the momentum variables are implicitly defined on a Euclidean space, with the canonical basis defined through the use of an identity mass matrix. Once again, it is not entirely clear how such a mass matrix could be chosen in a principled manner for different statistical models. Some rules of thumb have been suggested for tuning the mass matrix (35, 146) however these typically rely on some knowledge of the marginal distributions of the model parameters, which for most problems are of course unknown at the time of simulation. Exploratory runs of HMC, or indeed any other MCMC method, could be used to obtain estimates of such information, however such an approach is rather ad hoc and the exploratory simulations themselves may well have to be carefully tuned manually for individual problems. Even with prior knowledge regarding the marginal variances, different mass matrices may still be desirable for efficient sampling in the transient and stationary phases of the Markov chain. Given a fixed mass matrix the acceptance rate can of course be directly adjusted by choosing an appropriate integration step size, however we shall see in the following chapters how this does not necessarily translate into fast exploration of the target density.

2.6 Conclusions

Two separate ideas were developed around the 1950s for estimating the average properties of complex physical systems. Monte Carlo methods approached the problem by averaging over a large number of random samples, obtained for example by a Markov chain exploring the parameter space. Molecular Dynamics methods considered the evolution of a dynamical system based on differential equations describing its Hamiltonian mechanics and makes arguments based on ergodic theorems that the time average of such a system will approximately converge to the same value as the space average. We saw how a Langevin diffusion could also be used for this purpose. For physically realistic systems, Molecular Dynamics has the advantage of also being able to probe time dependent properties, such as the effect of perturbations to the system. We have seen how both of these approaches may usefully be applied to arbitrary statistical models,

and indeed the statistical properties of both methods may be improved by combining them in a single algorithm, Hamiltonian Monte Carlo.

The main problem associated with applying these methods to statistical models is that of setting the pre-conditioning or mass matrix. Indeed this is likely to be the reason that such methods have not been more enthusiastically adopted by the statistical community in general. In the next chapter we shall address this shortcoming by considering the use of geometric information to set the pre-conditioning matrix in MALA and the mass matrix in HMC.

While approaching this task we bear in mind the original problem of performing inference over statistical models based on systems of ODEs, where one of the main issues is the varying sensitivities of different parameter combinations on the model output, which directly affect the rate of change in probability mass of the target distribution with respect to these parameters. We will see that it is possible to take into account these sensitivities *automatically* and that by exploiting the local structure of the target density when proposing moves using these dynamical MCMC methods, we can drastically improve the overall statistical properties of our Markov chain, such as convergence and mixing. The local geometric structure of the target density may be conveniently expressed in terms of the Expected Fisher Information, which is obtained directly using the sensitivity equations of an ODE model. It turns out that the Fisher Information has some very attractive properties that provide us with an ideal starting point for developing a more general theory that allows us to integrate deeper geometric structure into a principled MCMC methodology.

3

Riemannian Manifold MCMC Methods

The dynamical sampling methods introduced in the previous chapter are implicitly defined on a simple Euclidean or vector space and there is in fact much more geometric information available to us; indeed statistical models have a natural geometric structure that is Riemannian in nature. We therefore generalise these sampling methods by defining them on a Riemannian manifold, such that proposal steps take into account the higher-order geometric information that is present. This results in more efficient sampling algorithms that can propose steps in the parameter space of a statistical model in which distance is measured in terms of the changes in probability mass, instead of the changes in the parameter values themselves. This chapter follows from work published with discussion in the Journal of the Royal Statistical Society: Series B (77).

3.1 Introduction

The idea of defining a distance between two parameterised probability distributions dates back more than 70 years to Rao (161) and Jeffreys (98). Much work has been done since then elucidating the relationship between statistics and Riemannian geometry, in particular examining geometric concepts such as distance, curvature and geodesics on statistical manifolds, within a field that has become known as Information Geometry (8). In the immediate 30 years after this relationship between statistics and differential geometry had been established, relatively little progress was made in

this area and it wasn't until Efron established a result characterising the curvature of exponential families (58) that there was a revival of interest. Much of the early work in this field in the 1970s and 1980s was primarily theoretical, with the development of geometries to describe higher order asymptotic theory, characterisation of exponential families and inference in nonlinear regression (103). Over the last 20 years there has been a second revival due to widely available high-speed computing, and differential geometry has subsequently had a much bigger impact on practical applications with more emphasis being placed on computational aspects. It has been applied to many other fields including econometrics (131), computer vision (136, 153), and machine learning (92, 121). Many of the required computations however can still be computationally expensive, and active areas of research often have to focus on how to implement these ideas in a computationally feasible manner. Calculating “straight” lines or geodesics between two points in a Riemannian geometry can be very useful, for example, in providing a natural method of interpolation for objects that can be represented as a Riemannian manifold, such as positive semi-definite matrices (153), however the balance between computational expense and gains in efficiency must be carefully weighed when considering this geometric approach.

3.1.1 Why Consider a Riemannian Geometry?

Our motivation for looking at the geometry of posterior distributions began with investigating Bayesian inference over the statistical models based on systems of ODEs that we can use to describe biological systems. In previous work (26, 27) it was observed that equal perturbations in different model parameters could have very different effects on the output of the ODE model, and hence on the resulting probability. Such differences are due to the strong nonlinearities that are often present in such statistical models, and indeed the nonlinearities are often necessary to accurately describe useful features of biological systems such as robustness; biochemical systems in plants, for example, need to cope with a wide variety of inputs, such as varying temperatures and levels of light, and yet still be able to function properly in terms of correctly regulating the levels of various proteins, which are modelled as outputs of the system. There are therefore certain parameters that have large effects on the output of such models when subject to small perturbations, and other parameters that can have very little effect even when perturbed by much larger amounts. When we perform Bayesian inference over such

systems, we want to find sets of parameters that have high probability of accurately characterising the data given a particular model; from this perspective it makes sense to define similarity between sets of parameters in terms of a distance based on the output rather than the input of the system.

Sensitivity analysis of dynamical systems frequently forms an important part of general investigation in systems biology and often such sensitivity analysis is performed by looking at some second order statistics of system perturbations, for example as described by the Expected Fisher Information. Somewhat conveniently, it is exactly this quantity that Rao (1961) discovered plays a useful role in defining distances between probability densities; indeed he showed that the use of the Expected Fisher Information as a metric endows a set of parameterised probability density functions with a Riemannian geometry. There is therefore a natural link between the sensitivity analysis of a statistical model and the definition of distance between model parameters in terms of a Riemannian geometry.

First order geometric information is already commonly employed in many optimisation methods and MCMC sampling algorithms. In some instances, their use may drastically speed up the convergence to a local maximum/minimum point or stationary distribution, respectively, however in other cases such algorithms exhibit very slow convergence. This happens when the gradients are not isotropic in magnitude (6); gradients may vary greatly in different directions and the rate of exploration of a parameter space may in addition be dependent on the problem-specific choice of parameterisation. This may often be seen most clearly in the proposed paths using Hamiltonian Monte Carlo, an example of which is given in Figure 3.1.

Methods using the standard gradient implicitly assume that the gradient in each direction is approximately constant over a small distance, when in fact these gradients may rapidly change over short distances. In Figure 3.1 the chain starts on the top of an elongated probability distribution. If we consider the standard gradient, then the steepest ascent is along the x-axis, rather than diagonally downwards towards the maximum point. When we consider the change in this gradient however, we see that it changes more rapidly along the x-axis than along the direction pointing toward the maximum. Intuitively, this means that moving “greedily” in the steepest direction may mean that we more rapidly reach a point at which we are moving down a gradient rather than up. Using second order geometric information therefore gives us a much

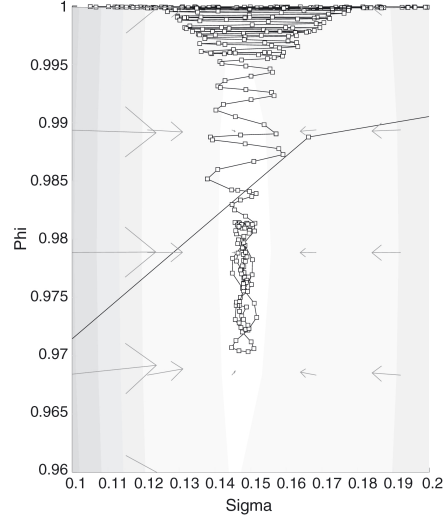


Figure 3.1: Slow convergence of a Markov chain using Hamiltonian Monte Carlo - This example is from the stochastic volatility model investigated in Chapter 4. It demonstrates HMC implicitly defined on a Euclidean space and slowly converging to a non-isotropic stationary distribution.

better idea of the direction in which we should be exploring, and in the optimisation literature this is often known as using the natural gradient (6); in other words, we wish to follow the steepest gradient *relative to the local geometry at the current point*, and this is intimately linked with the idea of proposing shortest paths across a Riemannian manifold.

3.1.2 A Quick Reminder of Euclidean Geometry

Let us first remind ourselves of some properties of Euclidean geometry before examining the merits of defining distance using a different geometry that is local instead of global. Many algorithms are implicitly defined on a vector space with a Euclidean geometry. Let us consider the n -dimensional real space \mathbb{R}^n as an example. Imposing a Euclidean geometry means employing a measure of distance between 2 points θ and $\theta + \epsilon\theta$ that is given by

$$D_E(\boldsymbol{\theta}, \boldsymbol{\theta} + \epsilon\boldsymbol{\theta}) = \sqrt{\sum_{i=1}^D \epsilon^2 \theta_i^2} \quad (3.1)$$

$$= \sqrt{\epsilon\boldsymbol{\theta}^T \cdot \epsilon\boldsymbol{\theta}} = \|\epsilon\boldsymbol{\theta}\| \quad (3.2)$$

which is the Euclidean norm. The same is true of standard MCMC algorithms; we implicitly use this canonical coordinate system to propose moves in the parameter space without considering in advance the likely changes in probability mass. We can see what this means in practice by thinking about how we measure distance on a map.

Maps often represent areas of the Earth as if they were flat and we use a distance measure that is isotropic regardless of the geographical landscape. According to the map representation, the distance between any two points is the same whether the terrain is a gentle hill or a steep mountain. It is obvious however that the distance travelled across the actual terrain depends not only on the 2-dimensional Euclidean coordinates but also on any changes in height, which we can describe using some function of the coordinates. This analogy transfers to the case of thinking about statistical models with the coordinates now representing the model parameter values and the height representing the probability mass at each point. In other words, a better measure of the distance between two points in a parameter space may have a more complicated structure defined in terms of the difference in some function of the parameters, rather than directly based on differences in the parameter values themselves.

As a simple example we can consider the distance between univariate normal distributions given a single data point $x = 0$, where the coordinates of our space are given by (μ, σ) , the mean and standard deviation, respectively. Let us consider a measure of dissimilarity between two points (μ_1, σ_1) and (μ_2, σ_2) in terms of the difference in likelihood of this simple statistical model. Calculating the likelihoods, we see that the difference between the points $(1, 1)$ and $(2, 1)$ is 0.188; this is much larger than the difference between $(1, 10)$ and $(2, 10)$, which is just 0.0006. On the other hand, using the standard Euclidean distance between parameter values we might come to the different opinion that both pairs of probability distributions are equally distant from one another, since $D_E = 1$ for both pairs.

The MALA method from Chapter 2 is implicitly defined with a Euclidean geometry; the standard gradient is used and the Brownian motion is sampled from an isotropic

3.2 An Introduction to Riemannian Geometry

Gaussian distribution. If we consider the pre-conditioned MALA, we see that this pre-multiplication is just a linear mapping and is therefore equivalent to exploring an inner product space, i.e. a vector space in which the linear mapping defines a change of basis. Similarly, the HMC approach is also defined with a Euclidean geometry if the mass matrix is equal to the identity matrix. Again, the mass matrix represents a linear mapping that defines the basis for the vector space of momentum variables \mathbf{p} , and so the basis for the parameters is given by the inverse mass matrix via the relation $\mathbf{v} = \mathbf{M}^{-1}\mathbf{p}$. By employing a fixed global metric in this manner, we assume that the local geometry is the same in all areas of the parameter space and that a small change in a particular parameter has the same effect on the output of the model irrespective of the values of the other parameters.

The final important point we should note is that geometry is to a large extent subjective. We may *impose* whichever geometry we wish, although usually we would choose one that is helpful in describing the model in some way. From a statistical point of view, it also makes sense that we should employ a geometry that has a deeper statistical interpretation. Other factors may also influence our choice of geometry; certain geometries may be computationally more tractable than others, for example. We should keep this important notion of the subjectiveness of geometry in mind for the rest of this chapter.

3.2 An Introduction to Riemannian Geometry

Curved spaces, far from being a mathematical abstraction, regularly occur in the real world; a simple example being the surface of the earth embedded within the 3 dimensional ambient space we inhabit, or even the shortest path taken by a beam of light travelling nearby a massive object in space. Indeed this was described by Einstein in his theory of relativity for which he used the formalism of Riemannian geometry.

For many statistical applications we do not need the full power of abstract differential geometry as presented in many modern mathematical texts. Indeed, also from a pedagogical point of view, too much abstraction makes it harder for cross-disciplinary researchers to pick up new concepts and ideas to use in their own work. This is particularly important in statistics, where statistical methodology may potentially be of practical use to researchers in a wide variety of fields, from biology to economics. In the

following introduction we therefore focus on the key concepts needed to work with such objects from a computational perspective, such that the ideas may be quickly picked up and easily implemented. There already exists a large selection of books available on the topic giving the complete mathematical details and exploring their intricacies more fully. Do Carmo offers a clear and very accessible introduction to the geometry of surfaces (31) and differential geometry (32), Marriott and Salmon provide an excellent account from a statistical perspective (131), and more detailed accounts are also available (34, 187, 205). The following introduction however should be sufficient for understanding the ideas presented in this thesis in the context of developing new MCMC methodology.

3.2.1 Differentiable Manifolds

We will first introduce the basics of Riemannian geometry from a more general perspective, and afterwards look at the Riemannian structure induced by a set of points that represent a family of probability distributions.

Informally, a manifold M is an n -dimensional space that is locally Euclidean; it is locally equivalent to \mathbb{R}^n via some smooth transformation. A more formal definition describes M as some set together with a collection of bijective functions (known as charts), which map overlapping sections of M to \mathbb{R}^n , such that there exists a map $f : M \mapsto \mathbb{R}^n$ such that f is a smooth continuous bijection. By using a combination of these charts, any path on M can be continuously mapped into \mathbb{R}^n . In this work however we assume that there exists a single continuous chart that maps the whole of M onto \mathbb{R}^n . Every point on the manifold can therefore be uniquely paired with a point in \mathbb{R}^n and vice versa. In general, a manifold is differentiable if f is infinitely differentiable and has continuous derivatives (32).

Given such a mathematical object, it is then possible to *impose* a geometry that defines a distance between any 2 points and consequently induces more global properties of the manifold itself, such as geodesics. From an MCMC perspective, we might sensibly regard locally similar points on a manifold to be those that have similar probability mass, however since our statistical model is parameterised in terms of the input parameters the key point to note is that changes in the parameter values do not necessarily correspond directly to changes in the probability of the model output; there may be a complex nonlinear relationship between the two, induced by the statistical model.

Being able to make appropriately sized steps for each parameter in terms of equal changes in probability could therefore plausibly lead to much improved exploration of the target distribution and the statistical properties of the samples drawn.

It is perhaps easiest to visualise a Riemannian manifold as an object embedded within a higher dimensional space, and indeed this will be a useful idea when we later consider a manifold from a statistical point of view. Riemannian manifolds can be fully described locally in terms of a metric tensor without reference to an ambient space, however it is more intuitive to learn the main concepts by thinking in terms of an embedding space. In fact, every n -dimensional manifold can be embedded in \mathbb{R}^m , where $m \geq n$, and this result is known as the Whitney embedding theorem. The surface of a sphere is a familiar example of a 2 dimensional manifold sitting in a 3 dimensional ambient or embedding space. A differentiable manifold becomes a Riemannian manifold when we assign a smoothly varying inner product to each point on the manifold, which in turn defines a tangent space at each point.

3.2.2 Tangent Spaces

At each point $\boldsymbol{\theta} \in \mathbb{R}^n$ on a Riemannian manifold M there exists a tangent space which we denote as $T_{\boldsymbol{\theta}}M$. We can think of this as a linear approximation to the Riemannian manifold at the point $\boldsymbol{\theta}$ and this is simply a standard vector space, whose origin is the current point on the manifold and whose vectors are tangent to this point. In terms of the sphere example previously, we may picture this vector space by imagining a sheet of stiff card balancing on top of the sphere at some point; all vectors in the space can be drawn on this sheet and the single point at which the card touches the sphere represents the origin of this vector space, as shown in Figure 3.2.

This tangent space exists independently of the coordinate system used and we can consider the basis of the tangent space on a manifold to be defined purely in terms of differential operators, which act on functions defining paths on the underlying manifold (53). The vector space $T_{\boldsymbol{\theta}}M$ is therefore spanned by the differential operators

$$\left[\frac{\partial}{\partial \theta_1}, \dots, \frac{\partial}{\partial \theta_n} \right] \tag{3.3}$$

which we may write in shorthand as $[\partial_1 \dots \partial_n]$. This leads to the interpretation of the basis vectors in the tangent space as directional derivatives; each differential operator

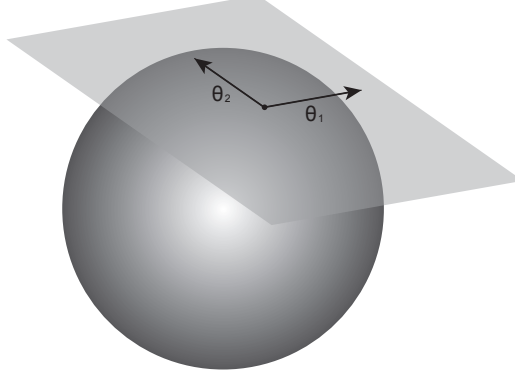


Figure 3.2: Representation of the tangent space on a sphere - The tangent space at a point on a sphere may be represented as all the vectors lying on a sheet of stiff card pinned to the sphere at a point that denotes the origin. The basis vectors of this tangent space are shown as θ_1 and θ_2 .

in the basis can act on a function on the manifold to give its rate of change in each direction.

3.2.3 Metric Tensors

The tangent space at each point θ arises when we equip our manifold with an inner product at each point, which we can use to measure distance and angles between vectors. This inner product is defined in terms of a metric tensor, G_θ , which defines the basis of each tangent space $T_\theta M$.

We now have the full definition that a Riemannian manifold is a differentiable manifold in which the tangent space at each point has an inner product defined via a metric tensor. A metric tensor is simply a generalisation of the dot product in Euclidean space; it is a function that takes 2 vectors, \mathbf{t}_1 and \mathbf{t}_2 , in the tangent space at some point θ on the manifold M , and produces a real-valued scalar $G_\theta(\mathbf{t}_1, \mathbf{t}_2)$,

$$G_\theta : T_\theta M \times T_\theta M \mapsto \mathbb{R} \tag{3.4}$$

This function can be used to define angles between two vectors and also the lengths of individual vectors. Each tangent vector $\mathbf{t}_1 \in T_\theta M$ at a point on the manifold $\theta \in M$ has a length $\|\mathbf{t}_1\| \in \mathbb{R}^+$, whose square is given by the inner product, such that

$$\|\mathbf{t}_1\|_{G_\theta}^2 = \langle \mathbf{t}_1, \mathbf{t}_1 \rangle_\theta = \mathbf{t}_1^T G_\theta \mathbf{t}_1 \quad (3.5)$$

This squared distance is known as the first fundamental form in Riemannian geometry (32) and is invariant to reparameterisations of the coordinates, as we shall see later when we consider the use of the Expected Fisher Information as a Riemannian metric tensor. We may then obtain a vector field on a manifold by assigning a tangent vector to each point such that this vector smoothly changes across the manifold, and this can be defined in terms of a smooth function from an n -dimensional manifold to the tangent space at each point, $\mathbf{f} : M \rightarrow T_\theta M$.

The Riemannian metric tensor is a symmetric, bilinear, positive definite function on M , such that

- $G_\theta(\mathbf{t}_1, \mathbf{t}_2) = G_\theta(\mathbf{t}_2, \mathbf{t}_1)$
- $G_\theta(\mathbf{t}_1 + \mathbf{t}_2, \mathbf{t}_3) = G_\theta(\mathbf{t}_1, \mathbf{t}_3) + G_\theta(\mathbf{t}_2, \mathbf{t}_3)$
- $G_\theta(\mathbf{t}_1, \mathbf{t}_1) > 0$

This metric tensor varies smoothly from point to point; for all vector fields on M , $G(\mathbf{t}_1, \mathbf{t}_2)(\theta) = G_\theta(\mathbf{t}_1, \mathbf{t}_2)$ is a smooth function of θ from M to \mathbb{R} , where $\mathbf{t}_1, \mathbf{t}_2 \in T_\theta M$, and we can therefore calculate derivatives of the metric tensor with respect to θ . It is also possible to calculate lengths of curves on the manifold (32). Let us assume we have some function that traces out a path on the manifold, $\theta(t) : \mathbb{R} \rightarrow M$, then the length of this curve on the manifold is given by

$$d_R(\theta(t)) = \int_{t_1}^{t_2} \sqrt{G_{\theta(t)} \left(\frac{d\theta}{dt}, \frac{d\theta}{dt} \right)} dt \quad (3.6)$$

$$= \int_{t_1}^{t_2} \left(\sum_{ij} G(\theta(t))_{ij} \frac{d\theta^i}{dt} \frac{d\theta^j}{dt} \right)^{\frac{1}{2}} dt \quad (3.7)$$

We can see that for a Euclidean geometry, when G is an identity matrix, Equation 3.6 simplifies to the standard equation for a line integral. Furthermore, if the metric tensor is constant for all values of θ then the Riemannian manifold is equivalent to a vector space with constant inner product.

3.2.4 Riemannian Manifolds from Statistical Models

We will now review how a family of probability distributions induced by a statistical model can be represented as a Riemannian manifold. It is useful to consider first an exponential family, which we shall present using a similar approach as (131, 144). In this section we make use of the more succinct Einstein notation, such that summations are denoted by adjacent upper and lower indices, for example

$$\theta^i s_i = \sum_{i=1}^I \theta_i s_i \quad (3.8)$$

Given a random variable \mathbf{x} this family of normalised probability distributions may be parameterised as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \exp(\theta^i s_i - A(\boldsymbol{\theta}))m(\mathbf{x}) \quad (3.9)$$

where $s(\mathbf{x})$ is a n -dimensional statistic, $\boldsymbol{\theta} \in \mathbb{R}^n$ the canonical parameter vector, and $m(\mathbf{x})$ is some non-negative function such that

$$\int \exp(\theta^i s_i) m(\mathbf{x}) d\mathbf{x} < \infty \quad (3.10)$$

The function $A(\boldsymbol{\theta})$ has the task of normalising the probability density function such that its total density is equal to one. The function $A(\boldsymbol{\theta})$ is automatically determined once the other functions and parameters have been specified.

One way of introducing a geometry on this family is to consider the log-likelihood,

$$\log [p(\mathbf{x}|\boldsymbol{\theta})] = \theta^i s_i - A(\boldsymbol{\theta}) + \log [m(\mathbf{x})] \quad (3.11)$$

$$\propto \theta^i s_i + \log [m(\mathbf{x})] \quad (3.12)$$

We note that the log-likelihood of this family of distributions is defined up to a constant of proportionality, and the key observation here is that this particular space of unnormalised log-densities has a very simple geometric structure; it forms an affine space. An affine space pair consists of a set and a vector space, (S, V) , along with a transformation

3.2 An Introduction to Riemannian Geometry

operation. Given a vector in V and a point in S , this transformation operation maps this pair to another point which also lies in S . An affine space is generally interpreted as a vector space without an origin point.

In the context of our unnormalised log exponential family, we can take our set to be \mathbb{R} , of which $\log [m(\mathbf{x}) \exp(\theta^i s_i)]$ is a member, and our vector space to be \mathbb{R}^n , in which our parameter vector $\boldsymbol{\theta}$ lies. We can then define our transformation to be

$$\begin{aligned} \log [m(\mathbf{x}) \exp(\theta^i s_i)] &= \log [m(\mathbf{x})] + \theta^i s_i \rightarrow \log [m(\mathbf{x})] + \theta^i s_i + \tilde{\theta}^i s_i \\ &= \log [m(\mathbf{x})] + (\theta^i + \tilde{\theta}^i) s_i \\ &= \log \left[m(\mathbf{x}) + \exp \left((\theta^i + \tilde{\theta}^i) s_i \right) \right] \end{aligned}$$

This transformation satisfies the properties of an affine space since it maps into the original set \mathbb{R} . We therefore have an affine space in which each point represents an unnormalised probability density in an exponential family. The choice of $m(\mathbf{x})$ determines the exponential family, and conversely every such exponential family forms an affine space.

We have seen that by working in log space, the unnormalised exponential families exhibit a rather simple geometry. We note again that this geometrical structure is rather arbitrary; we imposed this geometry just because it coincided with the simple structure that we noticed was naturally present in the log exponential family. We will see that this particular geometry is indeed very useful, however we bear in mind that other types of geometries may also be used to describe a family of probability distributions.

Let us now consider S as the set of all unnormalised log probability distributions, and imagine an unnormalised log exponential family that has some parameterisation different from the canonical parameters given in Equation 3.9. It is clear that this family will no longer necessarily have an affine structure, but rather may form a submanifold of the embedding space S as some kind of curved surface. Given such a representation we can still define what it means for a path on this manifold to be “straight” via the concept of a connection. In short, this dictates how vectors are mapped between tangent spaces and we shall discuss this in more detail in Section 3.2.6. It turns out that using this alternative description of “straightness” it is possible to characterise exponential families regardless of their parameterisation (58); we may identify them in terms of their

3.2 An Introduction to Riemannian Geometry

intrinsic qualities rather than the extrinsic parameter-dependent description, which hints at the power of using a differential geometric approach in statistics.

For now let us focus on some manifold, M , whose points are given by the parameters of an unnormalised density function, which we can consider as a log-likelihood of some statistical model given some data. At a particular point $\boldsymbol{\theta}$, the derivatives of the log-likelihood are tangent to the manifold and form a basis for the tangent space at $\boldsymbol{\theta}$, denoted by $T_{\boldsymbol{\theta}}M$. These tangent basis vectors are simply the score vectors at $\boldsymbol{\theta}$,

$$\nabla_{\boldsymbol{\theta}}\mathcal{L} = \left[\frac{\partial \mathcal{L}}{\partial \theta^1}, \dots, \frac{\partial \mathcal{L}}{\partial \theta^n} \right] \quad (3.13)$$

The tangent space is a linear approximation of the manifold at a given point and it has the same dimensionality. A natural inner product for this vector space is given by the covariance of the basis score vectors, since the covariance function satisfies the same properties, namely symmetry, bilinearity, and positive-definiteness. This inner product then turns out simply to be the Expected Fisher Information

$$G_{i,j} = \text{Cov} \left(\frac{\partial \mathcal{L}}{\partial \theta^i}, \frac{\partial \mathcal{L}}{\partial \theta^j} \right) \quad (3.14)$$

$$= E_{p(\mathbf{x}|\boldsymbol{\theta})} \left(\frac{\partial \mathcal{L}}{\partial \theta^i} \frac{\partial \mathcal{L}}{\partial \theta^j} \right) \quad (3.15)$$

which follows from the fact that the expectation of the score is zero,

$$E_{p(\mathbf{x}|\boldsymbol{\theta})} \left(\frac{\partial \mathcal{L}}{\partial \theta^i} \right) = \int \frac{1}{p(\mathbf{x}|\boldsymbol{\theta})} \frac{\partial}{\partial \theta^i} p(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \quad (3.16)$$

$$= \frac{\partial}{\partial \theta^i} \int p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \quad (3.17)$$

$$= 0 \quad (3.18)$$

The Expected Fisher Information can also be expressed in terms of second partial derivatives, which may be easier to compute for certain problems. This can be obtained by considering the expectation of the score function,

$$E_{p(\mathbf{x}|\boldsymbol{\theta})} \left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \right) = \int \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \quad (3.19)$$

$$= \int \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^2} p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} + \int \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \frac{\partial p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} d\mathbf{x} \quad (3.20)$$

$$= \int \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^2} p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} + \int \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}^T \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \quad (3.21)$$

$$= E_{p(\mathbf{x}|\boldsymbol{\theta})} \left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}^T \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \right) + E_{p(\mathbf{x}|\boldsymbol{\theta})} \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^2} \right) \quad (3.22)$$

$$= 0 \quad (3.23)$$

where we have used integration by parts and noted that the expectation of the score function is equal to zero. It then follows straightforwardly that

$$E_{p(\mathbf{x}|\boldsymbol{\theta})} \left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}^T \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \right) = -E_{p(\mathbf{x}|\boldsymbol{\theta})} \left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^2} \right) \quad (3.24)$$

Rao (161) and Jeffreys (99) showed that the Expected Fisher Information satisfies all the requirements for a metric tensor, and thus provides the extra structure needed to turn our differentiable manifold into a Riemannian manifold. In (10) the authors give the analytic expressions for geodesic distance between parameter values using this metric for a variety of families of probability distributions, although in general geodesic distances are analytically intractable. We shall see that while we need not necessarily choose to use the Expected Fisher Information, it does have some useful properties that may justify its choice.

Finally we note that although we have until now only considered exponential families, many other families of parameterised probability density functions can be considered as Riemannian manifolds provided they satisfy the following regularity conditions (8),

- All members have common support
- The basis vectors $\frac{\partial \mathcal{L}}{\partial \theta^i}$ are linearly independent
- Higher order moments of $\frac{\partial \mathcal{L}}{\partial \theta^i}$ exist
- Integration and differentiation are exchangeable with respect to parameters $\boldsymbol{\theta}$

It has been shown that all of these properties are indeed satisfied by exponential families (8).

3.2.5 Choosing a Metric

The Expected Fisher Information can often be convenient to use as a metric tensor. It arises naturally by considering the covariance of vectors tangent to a manifold, its inverse gives the best attainable asymptotic performance of any unbiased estimator (161), it is invariant to reparameterisations of the data \mathbf{x} (33, 40), and it has a natural interpretation in many areas of science (66).

We will now demonstrate the invariance of the first fundamental form when using the Expected Fisher Information specifically, remembering that this property more generally applies to all Riemannian metric tensors. The Expected Fisher Information defines a metric over a statistical model such that distances on the manifold of model parameters are invariant to reparameterisations. Let us first consider two statistical models, where one model is a reparameterisation of the other model, such that

$$\log p(\mathbf{x}|\phi) = \log p(\mathbf{x}|\boldsymbol{\theta}(\phi)) \quad (3.25)$$

If we take derivatives of each side with respect to ϕ it follows that

$$\frac{\partial \log p(\mathbf{x}|\phi)}{\partial \phi} = \sum_k \frac{\partial \log p(\mathbf{x}|\boldsymbol{\theta}(\phi))}{\partial \theta_k} \frac{\partial \theta_k}{\partial \phi} \quad (3.26)$$

The Expected Fisher Information then follows as

$$G(\phi) = \left(\frac{\partial \boldsymbol{\theta}}{\partial \phi} \right)^T G(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\theta}}{\partial \phi} \right) \quad (3.27)$$

and so under the new coordinates the Expected Fisher Information is just a transformation of original using the Jacobian of the reparameterisation. We can observe how this affects the definition of the inner product by considering two paths on the manifold, $\gamma(t)$ and $\tilde{\gamma}(t) = \phi(\gamma(t))$. Comparing the inner products of the tangent vectors given by the derivatives of our functions with respect to t , we see that

$$\begin{aligned}
\left(\frac{d\tilde{\gamma}(t)}{dt}\right)^T G(\phi) \left(\frac{d\tilde{\gamma}(t)}{dt}\right) &= \left(\frac{\partial\phi}{\partial\theta} \frac{d\gamma(t)}{dt}\right)^T \left(\frac{\partial\theta}{\partial\phi}\right)^T G(\theta) \left(\frac{\partial\theta}{\partial\phi}\right) \left(\frac{\partial\phi}{\partial\theta} \frac{d\gamma(t)}{dt}\right) \\
&= \left(\frac{d\gamma(t)}{dt}\right)^T \left(\frac{\partial\theta}{\partial\phi} \frac{\partial\phi}{\partial\theta}\right)^T G(\theta) \left(\frac{\partial\phi}{\partial\theta} \frac{\partial\theta}{\partial\phi}\right) \left(\frac{d\gamma(t)}{dt}\right) \\
&= \left(\frac{d\gamma(t)}{dt}\right)^T G(\theta) \left(\frac{d\gamma(t)}{dt}\right) \tag{3.28}
\end{aligned}$$

The inner product is therefore coordinate-independent on a Riemannian manifold, such that distance does not depend on the parameterisation of the statistical model. All Riemannian metric tensors transform covariantly, in a similar manner to the Expected Fisher Information. As a result the first fundamental form, defined as the squared distance, is invariant in any Riemannian geometry.

Although the Expected Fisher Information has some useful properties, it has perhaps one major disadvantage; for certain models it can be computationally intractable. If we insist on using it in such cases, then we must resort to approximation or estimation of this metric tensor (49, 186). An alternative is to note that we are not tied to the Expected Fisher Information and that we may in fact use any other metric tensor that satisfies the necessary conditions.

One useful method of obtaining new metrics is to derive them from other functions. In Euclidean space, distances between probability distributions P can be measured using functions that satisfy the following conditions

- Positive definiteness: $\forall P_1, P_2, d(P_1, P_2) > 0$
- Symmetry: $d(P_1, P_2) = d(P_2, P_1)$
- Triangle inequality: $\forall P_1, P_2, P_3, d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2)$

Since Riemannian manifolds are locally identifiable with Euclidean space, we require the same properties when defining a metric. Divergence functions also provide a measure of dissimilarity between probability distributions, indeed these types of functions were introduced around the same time that Rao investigated Riemannian distance between probability distributions. Such divergence functions have the property of being positive definite, however they do not define proper Euclidean distances as they do not satisfy

3.2 An Introduction to Riemannian Geometry

the symmetry and triangle inequality properties. One such function is the Kullback-Leibler divergence (115) between two probability distributions, which is defined as

$$KL(P_1||P_2) = \int P_1 \log \frac{P_1}{P_2} d\mathbf{x} \quad (3.29)$$

$$= E_{P_1} \left[\log \frac{P_1}{P_2} \right] \quad (3.30)$$

By considering a 2nd order Taylor approximation of the KL divergence between two nearby probability distributions $P_{\boldsymbol{\theta}_1}$ and $P_{\boldsymbol{\theta}_1+\epsilon\boldsymbol{\theta}}$ we obtain the Expected Fisher Information. Letting $u(\epsilon) = KL(P_{\boldsymbol{\theta}_1+\epsilon\boldsymbol{\theta}}||P_{\boldsymbol{\theta}_1})$, the 2nd order Taylor expansion is

$$u(\epsilon) \approx u(0) + \epsilon u'(0) + \frac{1}{2}\epsilon^2 u''(0) + O(\epsilon^3) \quad (3.31)$$

Since $u(0) = u'(0) = 0$, we are left with only the 2nd order term.

$$u(\epsilon) \approx \frac{\epsilon^2}{2} \sum_{ij} \theta^i \left[\int P_{\boldsymbol{\theta}_1} \left(\frac{1}{P_{\boldsymbol{\theta}_1}} \frac{\partial P_{\boldsymbol{\theta}_1}}{\partial \theta^i} \frac{1}{P_{\boldsymbol{\theta}_1}} \frac{\partial P_{\boldsymbol{\theta}_1}}{\partial \theta^j} \right) d\mathbf{x} \right] \theta^j + O(\epsilon^3) \quad (3.32)$$

$$= \frac{\epsilon^2}{2} \sum_{ij} \theta^i G_{ij}(\boldsymbol{\theta}_1) \theta^j + O(\epsilon^3) \quad (3.33)$$

$$= \frac{\epsilon^2}{2} \boldsymbol{\theta}^T G(\boldsymbol{\theta}_1) \boldsymbol{\theta} + O(\epsilon^3) \quad (3.34)$$

where G is simply the Expected Fisher Information. In fact metrics may also be derived from other types of divergences. Amari (7) showed that a 2nd order approximation of any divergence function gives a quadratic form that may be used as a metric,

$$D(P_{\boldsymbol{\theta}_1+d\boldsymbol{\theta}}||P_{\boldsymbol{\theta}_1}) \approx \frac{1}{2} \sum_{ij} G(\boldsymbol{\theta}_1)_{ij} d\theta^i d\theta^j \quad (3.35)$$

In particular the f-divergences introduced by Csiszar (45) all induce a unique Riemannian metric given by the Expected Fisher Information.

Another approach to deriving metrics was introduced by Burbea and Rao (23, 24) who showed that a large class of metrics may be obtained by considering ϕ -entropy functionals, which are based on convex functions. The ϕ -order entropy is defined as

$$H_\phi(P) \equiv - \int_X \phi(P) d\mathbf{x} \quad (3.36)$$

where P is some probability distribution in a parameterised family, and ϕ is any strictly positive, twice differentiable convex function. They showed that the squared Riemannian distance follows as

$$ds_\phi^2(\boldsymbol{\theta}) = -\Delta_{\boldsymbol{\theta}} H_\phi(P) \quad (3.37)$$

$$= \int \phi''(P) \left[\sum_{k=1}^D \frac{\partial P}{\partial \theta^k} d\theta^k \right]^2 \quad (3.38)$$

$$= \sum_{ij} G_{ij}^\phi d\theta^i d\theta^j \quad (3.39)$$

where the metric tensor is

$$G_{ij}^\phi = \int_{\mathbf{X}} \phi''(P) \frac{\partial P}{\partial \theta^i} \frac{\partial P}{\partial \theta^j} d\mathbf{x} \quad (3.40)$$

Choosing $\phi(P) = P \log P$ results once again in the Expected Fisher Information. We can choose other functions that result in different metrics for measuring distance on a Riemannian manifold of model parameters. One convenient set of such functions was introduced by Havrda and Charat (88)

$$\phi(P) = (\alpha - 1)^{-1} (P^\alpha - P), \quad \alpha \neq 1 \quad (3.41)$$

This function tends to $P \log P$ as α tends to 1, and produces the so-called α -order entropy metric tensors. By choosing $\alpha = 2$, the second derivative term of ϕ disappears and we obtain a metric with the simple form

$$G_{ij}^{\alpha=2} = 2 \int \frac{\partial P}{\partial \theta^i} \frac{\partial P}{\partial \theta^j} d\mathbf{x} \quad (3.42)$$

We see that this is particularly useful for mixtures of Gaussians, since in contrast to the Expected Fisher Information it is now analytically tractable. In addition the required derivatives of this metric tensor are also analytically tractable for a mixture

of Gaussians. Such ϕ -order entropy metric tensors have already been applied in the field of medical statistics, in particular for imaging applications such as shape matching (155).

Alternative versions of the Fisher Information are also available in the form of the Observed and Empirical Fisher Information. The Observed Fisher Information is simply the negative Hessian of the log target distribution and has the advantage of being more easily computable and based directly on the observed data. Unfortunately, if we consider the transformation properties of the negative Hessian we see that it does not always transform in a manner that preserves the inner product, and hence distances on the manifold may vary depending on the specific parameterisation used. Considering a transformation from parameters θ to ϕ , the Hessian transforms as

$$\frac{\partial \mathcal{L}}{\partial \phi \partial \phi} = \frac{\partial}{\partial \phi} \left[\frac{\partial \mathcal{L}}{\partial \theta} \frac{\partial \theta}{\partial \phi} \right] = \frac{\partial \theta^T}{\partial \phi} \frac{\partial \mathcal{L}}{\partial \theta \partial \theta} \frac{\partial \theta}{\partial \phi} + \frac{\partial \mathcal{L}}{\partial \theta} \frac{\partial^2 \theta}{\partial \phi \partial \phi} \quad (3.43)$$

and so only transforms correctly if the score vector is zero. The Expected Fisher Information on the other hand does transform correctly, since the expectation of the score vector is zero. In addition, the negative Hessian is not guaranteed to be positive definite and therefore does not constitute a proper metric tensor. Despite the potential numerical problems associated with this fact, it has been used as a method of accelerating optimisation, for example in Newton and quasi-Newton algorithms (64), and even in developing MCMC sampling schemes (158) with the use of somewhat ad hoc schemes to enforce positive definiteness. The Empirical Fisher Information is guaranteed to be positive definite, however small sample sizes can adversely affect the convergence of algorithms using it (64), and large sample sizes can result in greater computational cost.

3.2.6 Connecting Tangent Spaces

So far we have described the differential geometric structure at individual points on the manifold. There is just one more idea we need in order to be able to introduce ideas of dynamics on this manifold for developing MCMC methods, and the final piece is the concept of a *connection*. It is clear that if we want to develop MCMC schemes based on differential geometry we need to be able to move around this manifold and a connection tells us how to move vectors from the tangent space at one point on the

3.2 An Introduction to Riemannian Geometry

manifold, to the tangent space of a neighbouring point. In other words, a connection provides a mapping from the coordinate system of $T_{\theta}M$ to the coordinate system of $T_{\theta+\epsilon\theta}M$ (32).

Since a vector in $T_{\theta}M$ can be described as a linear combination of the basis vectors of $T_{\theta}M$, it suffices to look at how these basis vectors transform. In particular we want to know the rate of change of basis vector ∂_j as it moves in the direction of the basis vector ∂_i . Again this is simply the covariant or directional derivative of the basis vector. This operation results in another vector in tangent space, which we can again write as a linear combination of the basis vectors as follows

$$\nabla_{\partial_i}\partial_j = \Gamma_{ij}^m\partial_m \quad (3.44)$$

noting the implicit sum over the basis vectors using Einstein notation. Here we have used the Christoffel symbols as a shorthand. The Christoffel symbol Γ_{ij}^m denotes the coefficient of the m th basis vector, obtained by moving the j th basis vector in the i th direction. We can now see how a connection can be used to transport a vector from the tangent space $T_{\theta}M$ to $T_{\theta+d\theta}M$

$$v_{\theta+d\theta}^j = v_{\theta}^j + d\theta^i\Gamma_{ij}^m\partial_m \quad (3.45)$$

The vector $d\theta^i\Gamma_{ij}^m\partial_m$ gives the total change in the j th component of our vector v_{θ} , by summing over the effect of moving a distance θ^i in each direction ∂_i . This is visualised in Figure 3.3.

Once again there are many ways of mapping each tangent space to its neighbours, recalling that we may choose the geometry, with some choices being more suited to certain situations than others. The decisions we make about which geometry to employ can for instance be made in terms of the statistical interpretation of the geometry, or sometimes simply in terms of computational convenience.

Another use of a connection is that it can be used to define the notion of straightness on a manifold. *Geodesics* on a manifold are the equivalent of straight lines in Euclidean space (32). We define geodesics in terms of mapping vectors along paths on the manifold; if we take a vector in the tangent space of a point on a geodesic path and map this vector to the tangent space of any other point on this path using some connection,

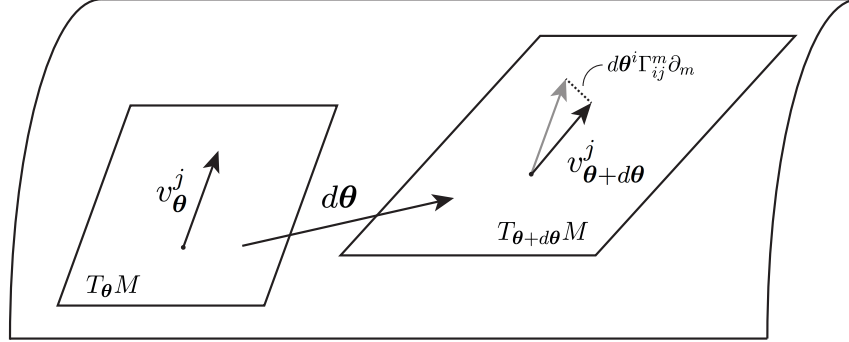


Figure 3.3: Representation of a connection in a Riemannian manifold - A connection in a Riemannian manifold provides a mapping of vectors from a tangent space to another nearby tangent space in terms of changes in the underlying basis vectors.

then this tangent vector will still point in the same direction with the same magnitude. It is immediately clear that different connections will induce different geodesic paths on a given manifold, since the connection defines how tangent vectors map across nearby tangent spaces. The condition for “straightness” on a manifold is therefore given by the equation

$$\nabla_{\dot{\theta}} \dot{\theta} = 0 \quad (3.46)$$

where we have a parameterised path $\theta(t)$ along the manifold with tangent vector $\dot{\theta}$, which is the directional derivative along the path of θ with respect to t . In this equation we impose the condition that the rate of change of the basis vectors of the tangent space along a geodesic must be zero.

It is interesting to note that in Euclidean space the straight lines always give the shortest route between two points, however the same is not always true for geodesics on a manifold. On any particular manifold, the geodesics do not necessarily coincide when using different connections. Given two points on a manifold, we could therefore obtain two different geodesics when using different mappings between tangent spaces, and these geodesic paths could have different lengths. This problem is resolved through the fundamental lemma of Riemannian geometry, which states that there is a unique

3.2 An Introduction to Riemannian Geometry

connection whose geodesics are locally¹ of shortest length. This is given by the Levi-Civita connection (32), which is sometimes also called the Riemannian connection or metric connection. The Levi-Civita connection is given by

$$\Gamma_{ij}^k = \frac{1}{2} G_{km}^{-1} (\partial_i G_{jm} + \partial_j G_{im} - \partial_m G_{ij}) \quad (3.47)$$

As we have seen, geodesics on a manifold can be described as the solution to Equation 3.46. This is more commonly written in component form, in terms of the chosen metric tensor and connection. If we define a parameterised path on the manifold as $\theta(t)$, then the geodesics are those curves that satisfy the 2nd order differential equation

$$\frac{\partial^2 \theta^m}{\partial t^2} + \Gamma_{ij}^m \frac{d\theta^i}{dt} \frac{d\theta^j}{dt} = 0 \quad (3.48)$$

This is the Euler-Lagrange equation and it is closely linked with a variational approach to mechanics in which this equation can be derived based on the principle of stationary action (83); this equation can also be more conveniently rewritten as Hamilton's equations and, given suitable initial conditions, it defines the path on a Riemannian manifold along which the total energy, given by the Hamiltonian, is constant. This provides a natural method for making proposals on a manifold for MCMC algorithms, and provides a means of obtaining a Hamiltonian Monte Carlo algorithm defined on a Riemannian manifold.

Given this method of following geodesics on a manifold, we can also choose the type of connection we wish to employ, through which the concept of “straightness” is defined. Although the Levi-Civita connection is a natural choice in the context of MCMC, since it defines geodesics that are also paths of minimum length between two points, there are a variety of other possible so-called non-metric or non-Riemannian connections we could choose to use instead. An important example is the a family of α -connections introduced by Cencov (33), Dawid (51, 52) and Amari (5). These connections are indexed by the real-valued parameter α , and coincide with the Levi-Civita connection when $\alpha = 0$. The α -connections are defined as

¹Consider for example the geodesics on a sphere, which are given by the great circles. We can travel round the geodesic in two possible directions, only one of which will generally be the shortest route. Geodesics can therefore only be described in terms of locally minimising distance.

3.2 An Introduction to Riemannian Geometry

$$\Gamma_{i,jk}^{\alpha} = \Gamma_{i,jk}^0 - \frac{\alpha}{2} T_{i,jk} \quad (3.49)$$

where T is the skewness tensor, defined as

$$T_{i,jk}(\alpha) = E_{p(\mathbf{x}|\boldsymbol{\theta})} \left[\frac{\partial \mathcal{L}}{\partial \theta^i} \frac{\partial \mathcal{L}}{\partial \theta^j} \frac{\partial \mathcal{L}}{\partial \theta^k} \right] \quad (3.50)$$

These connections have useful interpretations in different statistical contexts, particularly when considering the flatness of a manifold. A manifold can be considered flat when there exists a coordinate system under which the metric tensor is independent of the parameters (103); in particular exponential families form a flat manifold under the +1-connection, and families of mixtures form a flat manifold under the -1 -connection. The +1-connection is closely related to the work presented in Efron's seminal paper (58), in which he introduced the idea of statistical curvature based on this connection and uses it to quantify how well arbitrary probability distributions might be approximated using distributions from a curved exponential family. Such curvature is given by the Riemann curvature tensor and is an intrinsic measure independent of parameterisation, as opposed to the imbedding curvature which is a measure of curvature relative to the embedding space (see e.g. (32, 119)). It is also known that there is a natural duality between pairs of α -connections; if a manifold is flat using a $+\alpha$ -connection then it will also be flat using a $-\alpha$ -connection, although statistical interpretations of these properties are still the subject of current research (8). A statistical manifold can be defined as the triple (M, G, T) , where M is a manifold, G is a metric tensor, and T is a skewness tensor, and this was introduced by Lauritzen (119) as a means of unifying the various related geometries suggested previously (5, 14, 59). An alternative general framework which aims to unify this work is the preferred point geometry of Critchley (43, 44), which provides a natural way of describing an asymmetric geometry in which a chosen "preferred point" on the manifold plays a significant role. For example, if the data is known to have been generated from a parameterised family of distributions $p(\mathbf{x}|\boldsymbol{\theta})$ for some unknown value $\boldsymbol{\theta} = \boldsymbol{\Phi}$ then this information can be encoded using a preferred point metric, which is defined as

$$G_{ij}^{\boldsymbol{\Phi}}(\boldsymbol{\theta}) = \text{Cov}_{p(\mathbf{x}|\boldsymbol{\Phi})} \left(\frac{\partial \mathcal{L}}{\partial \theta^i} \frac{\partial \mathcal{L}}{\partial \theta^j} \right) \quad (3.51)$$

The metric that this geometry induces however is not a valid Riemannian metric in general, as it only transforms covariantly when the preferred point is equal to the current parameters, in which case it coincides with the Expected Fisher Information.

There is clearly much greater structure available to us by considering the more detailed, higher order geometries described in this general framework, however from a practical point of view there is an important trade-off between complexity and the additional benefit such complexity brings. Such an approach opens up the possibility of introducing particular geometries suited to specific statistical models, although in this work we decide to focus on developing geometric MCMC methods that are as widely applicable as possible; with this aim in mind it seems that the standard Levi-Civita (distance minimising) connection is most appropriate for this general and practical purpose.

3.3 Riemannian Manifold MALA

We have seen how a family of probability distributions has a natural representation as a Riemannian manifold, and we now return to the problem of sampling the parameters of such families. We first consider a manifold version of the MALA sampling algorithm, which we saw made proposal steps based on a stochastic differential equation defining a Langevin diffusion. It turns out we can also define such a diffusion on a Riemannian manifold, and so in a similar manner we can derive a sampling algorithm that takes the underlying geometric structure into account when making proposals.

The manifold version of this Langevin diffusion equation is well known and can be used to describe, for example, heat flow across a surface. It is based on the Laplace-Beltrami operator (39, 105), which simply measures the divergence of a vector field on a manifold. The stochastic differential equation defining the Langevin diffusion on a manifold is

$$d\boldsymbol{\theta}(t) = \frac{1}{2} \tilde{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(t)) dt + d\mathbf{b}(t) \quad (3.52)$$

where the natural gradient (6) is $\tilde{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(t)) = G^{-1}(\boldsymbol{\theta}(t)) \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}(t))$ and the Brownian motion on the Riemannian manifold is defined as

$$d\tilde{\mathbf{b}}_i(t) = |G(\boldsymbol{\theta}(t))|^{-\frac{1}{2}} \sum_{j=1}^D \frac{\partial}{\partial \theta_j} (G^{-1}(\boldsymbol{\theta}(t))_{ij} |G(\boldsymbol{\theta}(t))|^{\frac{1}{2}}) dt + \left(\sqrt{G^{-1}(\boldsymbol{\theta}(t))} d\mathbf{b}(t) \right)_i \quad (3.53)$$

In order to understand how this equation arises we start with a couple of definitions. Firstly, the gradient of a function on manifold, $f : \mathbf{M} \rightarrow \mathbb{R}$, is defined as the vector field

$$(\text{grad} f)_i = \sum_j G_{ij}^{-1} \frac{\partial f}{\partial \theta_j} \quad (3.54)$$

We can also consider this as the gradient of the function translated into the tangent space (a vector space) at the current point by a linear transformation using the basis vectors defined by the metric tensor. Secondly, the divergence operator measures the rate of change of a volume element as it moves along a vector field. Just as we considered the idea of volume preservation of symplectic integrators in Chapter 2, so we can also define an infinitesimal volume element on our manifold. On a Riemannian manifold there is a natural volume form defined as

$$d\mathbf{v} = \sqrt{\det(\mathbf{G})} d\boldsymbol{\theta} \quad (3.55)$$

and this is defined such that it is invariant under reparameterisations of the coordinates; indeed it was this fact that lead Jeffreys to propose his eponymous prior (99). We see that if the metric tensor is given by an identity matrix, then the volume element is the standard Euclidean volume. The divergence operator measures the rate of change of this volume element along a vector field and is defined as

$$\text{div} \mathbf{V} = \frac{1}{\sqrt{\det(\mathbf{G})}} \sum_i \frac{\partial}{\partial \theta_i} \left(\sqrt{\det(\mathbf{G})} \mathbf{V}_i \right) \quad (3.56)$$

where \mathbf{V} is a vector field. Finally, the Laplace-Beltrami operator is defined using both the gradient and the divergence

$$\Delta = \text{div} \circ \text{grad} \quad (3.57)$$

and this acts on functions defined on the manifold such that

$$\Delta_i f = \frac{1}{\sqrt{\det(\mathbf{G})}} \sum_i \frac{\partial}{\partial \theta_i} \left(\sum_j G_{ij}^{-1} \sqrt{\det(\mathbf{G})} \frac{\partial f}{\partial \theta_j} \right) \quad (3.58)$$

The stochastic differential equation defining the Langevin diffusion over a Riemannian manifold is therefore made up of a drift term, which is a movement based on the natural gradient of a function on the manifold, and a diffusion term which is defined based on the Laplace-Beltrami operator. Indeed the SDE for Brownian motion on a manifold is defined using the Laplace-Beltrami operator as its generator (93). Once again we see that when the metric tensor is given by an identity matrix, then the Laplace-Beltrami operator reduces to the standard Laplacian operator, which is simply the divergence of the gradient on a Euclidean space.

Given the geometric structure for a statistical model, a Langevin diffusion with invariant measure $p(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^D$, can be defined directly upon the Riemannian manifold with metric tensor $G(\boldsymbol{\theta})$ (105). Clearly in a Euclidean space where the metric tensor is an identity matrix then Equation 3.52 reduces to the standard form of SDE that we used to describe a Langevin diffusion in Chapter 2. The first part of the right hand side of Equation 3.53 represents the 1st order terms of the Laplace-Beltrami operator and these relate to the local curvature of the manifold, reducing to zero if the metric is everywhere constant. The second term on the right hand side provides a position specific linear transformation of the Brownian motion $\mathbf{b}(t)$ based on the local metric.

Since the metric tensor is a function of the parameters $\boldsymbol{\theta}$, we can expand the expression for the i th component of the 1st order terms based on the Laplace-Beltrami operator by differentiating as follows

$$\Delta_i = \frac{1}{\sqrt{\det(G)}} \sum_j \frac{\partial}{\partial \theta_j} \left((G^{-1})_{ij} \sqrt{\det(G)} \right) \quad (3.59)$$

$$= |G|^{-\frac{1}{2}} \sum_j \frac{\partial}{\partial \theta_j} \left((G^{-1})_{ij} |G|^{\frac{1}{2}} \right) + |G|^{-\frac{1}{2}} \sum_j \left((G^{-1})_{ij} \frac{\partial}{\partial \theta_j} |G|^{\frac{1}{2}} \right) \quad (3.60)$$

$$= - \sum_j \left(G^{-1} \frac{\partial G}{\partial \theta_j} G^{-1} \right)_{ij} + |G|^{-\frac{1}{2}} \sum_j \left((G^{-1})_{ij} \frac{1}{2} |G|^{-\frac{1}{2}} \frac{\partial}{\partial \theta_j} |G| \right) \quad (3.61)$$

$$= - \sum_j \left(G^{-1} \frac{\partial G}{\partial \theta_j} G^{-1} \right)_{ij} + \frac{1}{2} |G|^{-1} \sum_j \left((G^{-1})_{ij} |G| \text{trace} \left(G^{-1} \frac{\partial G}{\partial \theta_j} \right) \right) \quad (3.62)$$

$$= - \sum_j \left(G^{-1} \frac{\partial G}{\partial \theta_j} G^{-1} \right)_{ij} + \frac{1}{2} \sum_j \left((G^{-1})_{ij} \text{trace} \left(G^{-1} \frac{\partial G}{\partial \theta_j} \right) \right) \quad (3.63)$$

where we have standard expressions for matrix calculus (156). Employing a first order Euler integrator, the discrete form of the stochastic differential equation 3.52 therefore follows as

$$\theta_i^{n+1} = \theta_i^n + \frac{\epsilon^2}{2} (G^{-1}(\theta^n) \nabla_{\theta} \mathcal{L}(\theta^n))_i - \epsilon^2 \sum_{j=1}^D \left(G^{-1}(\theta^n) \frac{\partial G(\theta^n)}{\partial \theta_j} G^{-1}(\theta^n) \right)_{ij} \quad (3.64)$$

$$+ \frac{\epsilon^2}{2} \sum_{j=1}^D (G^{-1}(\theta^n))_{ij} \text{Tr} \left(G^{-1}(\theta^n) \frac{\partial G(\theta^n)}{\partial \theta_j} \right) + \left(\epsilon \sqrt{G^{-1}(\theta^n)} \mathbf{z}^n \right)_i \quad (3.65)$$

$$= \boldsymbol{\mu}(\theta^n, \epsilon)_i + \left(\epsilon \sqrt{G^{-1}(\theta^n)} \mathbf{z}^n \right)_i \quad (3.66)$$

which defines a proposal mechanism with density $q(\theta^* | \theta^n) = \mathcal{N}(\theta^* | \boldsymbol{\mu}(\theta^n, \epsilon), \epsilon^2 G^{-1}(\theta^n))$ and acceptance probability $\min\{1, p(\theta^*)q(\theta^n | \theta^*)/p(\theta^n)q(\theta^* | \theta^n)\}$ to ensure convergence to the invariant density $p(\theta)$. Immediately it is clear that the proposal mechanism makes moves approximately along the D -dimensional manifold rather than the D -dimensional Euclidean space, and that these moves respect the curvature at each point of the manifold. Pseudo-code describing the full manifold MALA (mMALA) scheme is given by Algorithm 2.

It may be computationally expensive to calculate the 3rd order derivatives needed for working out the rate of change of the metric tensor, and so an obvious approximation is to assume these derivatives are zero for each step. In other words, for each step we can assume that the metric is locally constant. Of course even if the curvature of the manifold is not constant, this simplified proposal mechanism still defines a

Algorithm 2 Manifold MALA

```

1: Initialise current  $\theta$ 
2: for IterationNum = 1 to NumSamples do
3:   Sample  $\theta^{\text{new}}$  based on Current  $\theta$  according to first order discretisation
4:   Calculate current log-likelihood  $\mathcal{L}(\theta)$  and proposed log-likelihood  $\mathcal{L}(\theta^{\text{new}})$ 
5:   Calculate  $\log(p(\theta^{\text{new}}|\theta))$ ,  $\log(p(\theta|\theta^{\text{new}}))$ ,  $\log(\text{Prior}(\theta))$ ,  $\log(\text{Prior}(\theta^{\text{new}}))$ 
6:   LogRatio =  $\mathcal{L}(\theta^{\text{new}}) + \log(\text{Prior}(\theta^{\text{new}})) + \log(p(\theta|\theta^{\text{new}})) - \mathcal{L}(\theta) - \log(\text{Prior}(\theta)) - \log(p(\theta^{\text{new}}|\theta))$ 
7:   % Accept or reject according to Metropolis ratio
8:   if LogRatio > 0 or LogRatio > log(rand) then
9:     Set  $\theta = \theta^{\text{new}}$ 
10:  end if
11: end for

```

correct MCMC method which converges to the target measure, as we accept or reject moves using a Metropolis-Hastings ratio. This is equivalent to a position specific pre-conditioned MALA proposal, where the preconditioning is dependent on the current parameter values

$$\theta^{n+1} = \theta^n + \frac{\epsilon^2}{2} G^{-1}(\theta^n) \nabla_{\theta} \mathcal{L}(\theta^n) + \epsilon \sqrt{G^{-1}(\theta^n)} \mathbf{z}^n \quad (3.67)$$

For a manifold whose metric tensor is globally constant, this reduces further to the preconditioned MALA proposal we saw in the last chapter, where the preconditioning is effectively independent of the current parameter values. However, in contrast to before pre-conditioning no longer needs to be chosen arbitrarily, but rather is now informed by the geometry of the distribution we are exploring.

Intuitively we might expect such approximate proposal processes to be less efficient in converging to the stationary distribution, since it is no longer following the curvature of the target distribution precisely. We shall explore this further in the experimental evaluation in the next chapter. We now consider a simple example for the purposes of illustrating this geometric approach and gaining some insight into the mMALA method.

3.3.1 Illustrative Example: Parameters of a Gaussian Distribution

We return to the simple example from Chapter 2 of inferring the parameters of a Gaussian distribution. We again consider $N = 30$ observations drawn from a Gaussian distribution $\mathcal{N}(\mathbf{y}|\mu = 0, \sigma = 10)$. We shall investigate the differences in sampling the model parameters using both MALA and mMALA schemes. For the MALA we simply need the log-likelihood and its derivatives. For mMALA we also require the metric tensor, given by the Expected Fisher Information, and its derivatives. The second partial derivatives of the log-likelihood follow as

$$\frac{\partial^2 L}{\partial \mu^2} = -\frac{N}{\sigma^2} \quad (3.68)$$

$$\frac{\partial^2 L}{\partial \sigma^2} = \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{n=1}^N (y_n - \mu)^2 \quad (3.69)$$

$$\frac{\partial^2 L}{\partial \sigma \partial \mu} = \frac{\partial^2 L}{\partial \mu \partial \sigma} = -\frac{2}{\sigma^3} \sum_{n=1}^N (y_n - \mu) \quad (3.70)$$

and, noting that $E[(y_n - \mu)^2] = \sigma^2$, the metric tensor and its derivatives are simply

$$\mathbf{G} = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{2N}{\sigma^2} \end{bmatrix} \quad \frac{\partial \mathbf{G}}{\partial \mu} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \frac{\partial \mathbf{G}}{\partial \sigma} = \begin{bmatrix} -\frac{2N}{\sigma^3} & 0 \\ 0 & -\frac{4N}{\sigma^3} \end{bmatrix} \quad (3.71)$$

Starting at the point $[\mu = 0, \sigma = 10]$, we propose 50 steps using the MALA, mMALA and simplified mMALA samplers, with stepsizes 0.005, 1 and 1 respectively. For MALA, 0.005 was the largest stepsize with which proposed moves were accepted from the starting point. We see in Figure 3.4 how the scaling changes quickly using MALA; large moves are taken at first, but as the gradient decreases the moves become smaller and smaller. After 50 steps the chain is still some way from the mode, and most likely a different stepsize is necessary in different parts of the parameter space. The manifold methods perform far better, with both mMALA and simplified mMALA reaching the mode without any need to rescale the stepsize.

3.4 Riemannian Manifold Hamiltonian Monte Carlo

We can now define the Hamiltonian that we saw in Chapter 2 in a more general form on a Riemannian manifold and this allows us to obtain a Hamiltonian Monte

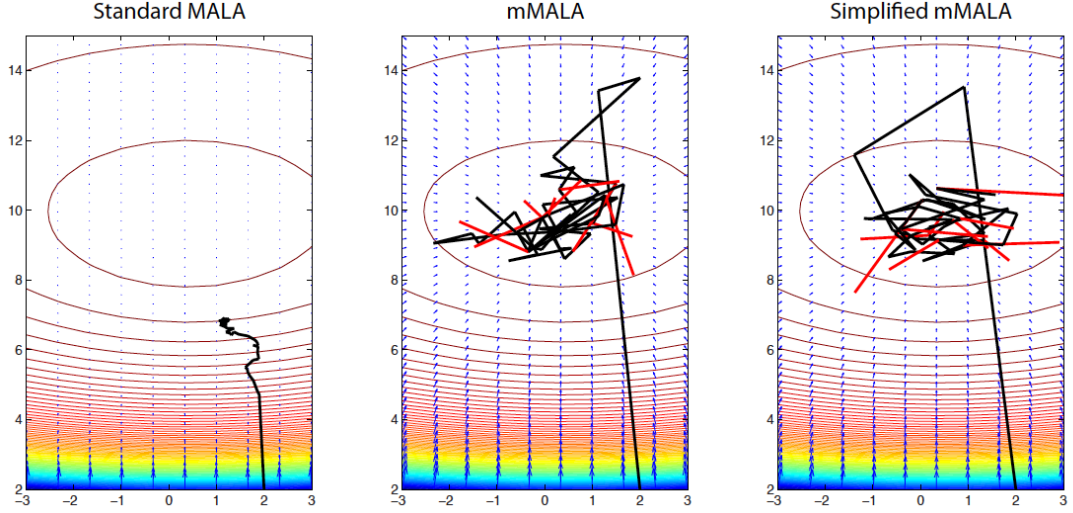


Figure 3.4: Comparison of MALA, mMALA and simplified mMALA samplers using a simple Gaussian model - Accepted moves are shown in black and rejected moves in red. Scaling issues dramatically affect convergence of MALA with a fixed stepsize.

Carlo method that exploits the local geometric structure induced by a statistical model. Zlochin and Baram (208) originally attempted to exploit this manifold structure within a Molecular Dynamics framework, however their use of a non-reversible and non-symplectic numerical integration method meant that samples were not drawn from the correct stationary distribution; indeed even using a reversible, symplectic integration method we still need to correct for discretisation errors using a Metropolis-Hastings step to guarantee a statistically correct MCMC procedure.

The definition of a Hamiltonian on a Riemannian manifold is straightforward. From Newtonian mechanics we know that $\mathbf{p} = \mathbf{M}\dot{\boldsymbol{\theta}}$, and so the squared distance of each tangent vector $\dot{\boldsymbol{\theta}}$ under some metric \mathbf{M} can be written in terms of the momentum

$$\|\dot{\boldsymbol{\theta}}\|_{\mathbf{M}}^2 = \dot{\boldsymbol{\theta}}^T \mathbf{M} \dot{\boldsymbol{\theta}} = \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} \quad (3.72)$$

In a more general form, we can consider the case where our parameters $\boldsymbol{\theta}$ define a Riemannian manifold with metric tensor $G(\boldsymbol{\theta})$, which is induced by some statistical model using for example the Expected Fisher Information. This now defines a position specific squared distance such that

$$\|\dot{\boldsymbol{\theta}}\|_{G(\boldsymbol{\theta})}^2 = \dot{\boldsymbol{\theta}}^T G(\boldsymbol{\theta}) \dot{\boldsymbol{\theta}} = \mathbf{p}^T G(\boldsymbol{\theta})^{-1} \mathbf{p} \quad (3.73)$$

We can define the kinetic energy term via a position dependent inverse metric tensor (30). We now note that if we want to interpret this as the log-density of a Gaussian distribution, as we did with the original HMC method, we must add a normalising constant since this changes with the metric from point to point. We can therefore define a Hamiltonian on the Riemannian manifold as

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2} \log((2\pi)^D |G(\boldsymbol{\theta})|) + \frac{1}{2} \mathbf{p}^T G(\boldsymbol{\theta})^{-1} \mathbf{p} \quad (3.74)$$

so that the Hamiltonian can be considered as the negative joint log-likelihood of the target density and the auxiliary variable \mathbf{p} , since

$$\exp(-H(\boldsymbol{\theta}, \mathbf{p})) = p(\boldsymbol{\theta}, \mathbf{p}) = p(\boldsymbol{\theta})p(\mathbf{p}|\boldsymbol{\theta}) \quad (3.75)$$

The desired target density is once again simply the marginal density

$$\begin{aligned} p(\boldsymbol{\theta}) \propto \int \exp(-H(\boldsymbol{\theta}, \mathbf{p})) d\mathbf{p} &= \frac{\exp\{\mathcal{L}(\boldsymbol{\theta})\}}{\sqrt{(2\pi)^D |G(\boldsymbol{\theta})|}} \int \exp\left\{-\frac{1}{2} \mathbf{p}^T G(\boldsymbol{\theta})^{-1} \mathbf{p}\right\} d\mathbf{p} \\ &= \exp\{\mathcal{L}(\boldsymbol{\theta})\} \end{aligned}$$

Unlike in the previous case for HMC, this joint density is no longer factorisable and therefore the log-likelihood does not correspond to a separable Hamiltonian. The conditional distribution for momentum values given parameter values is a zero-mean Gaussian with the point specific metric tensor acting as the covariance matrix

$$p(\mathbf{p}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, G(\boldsymbol{\theta})) \quad (3.76)$$

which like mMALA should resolve the problems associated with tuning the mass matrix in HMC. We will investigate this further for a variety statistical model examples in Chapter 4. The dynamics are once again defined by Hamilton's equations as

$$\begin{aligned}\frac{d\theta_i}{d\tau} &= \frac{\partial H}{\partial p_i} = (G(\boldsymbol{\theta})^{-1}\mathbf{p})_i \\ \frac{dp_i}{d\tau} &= -\frac{\partial H}{\partial \theta_i} = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} - \frac{1}{2}\text{Tr} \left[G(\boldsymbol{\theta})^{-1} \frac{\partial G(\boldsymbol{\theta})}{\partial \theta_i} \right] + \frac{1}{2}\mathbf{p}^T G(\boldsymbol{\theta})^{-1} \frac{\partial G(\boldsymbol{\theta})}{\partial \theta_i} G(\boldsymbol{\theta})^{-1} \mathbf{p}\end{aligned}$$

The Hamiltonian dynamics on the manifold are simulated by solving these first order differential equations, which are equivalent to calculating the geodesic path given by the Euler-Lagrange equation (Equation 3.48) earlier in this chapter (83). These continuous-time Hamiltonian equations are energy preserving, volume preserving and time reversible, as we proved in Chapter 2. When we discretise the dynamics, however, numerical integration is no longer quite so straightforward as it was for a separable Hamiltonian. Naively employing the discrete Störmer-Verlet Leapfrog integrator (Equation 2.62) it is clear that for a finite step size ϵ the mappings for $\boldsymbol{\theta}$ and \mathbf{p} are no longer reversible, since the metric tensor is position dependent, and in general $G(\boldsymbol{\theta}(\tau)) \neq G(\boldsymbol{\theta}(\tau + \epsilon))$; as a result, measures of distance will be different at different points and reverse steps will not move back to exactly the original position. Proposals generated using this integrator will therefore not satisfy detailed balance in a Hamiltonian Monte Carlo scheme.

We require a time-reversible, volume preserving numerical integrator for solving this *non-separable* Hamiltonian to ensure a correct MCMC algorithm. Such a second-order integrator can be formed by the composition of the first-order symplectic Euler integrators we saw in Chapter 2. This is the Generalised Leapfrog integration scheme and we recap that it follows as

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^n - \frac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^n, \mathbf{p}^{n+\frac{1}{2}}) \quad (3.77)$$

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \frac{\epsilon}{2} \left[\nabla_{\mathbf{p}} H(\boldsymbol{\theta}^n, \mathbf{p}^{n+\frac{1}{2}}) + \nabla_{\mathbf{p}} H(\boldsymbol{\theta}^{n+1}, \mathbf{p}^{n+\frac{1}{2}}) \right] \quad (3.78)$$

$$\mathbf{p}_{n+1} = \mathbf{p}^{n+\frac{1}{2}} - \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}^{n+1}, \mathbf{p}^{n+\frac{1}{2}}) \quad (3.79)$$

If the Hamiltonian is separable then the Generalised Leapfrog reduces to the standard Störmer-Verlet Leapfrog scheme that is commonly used in HMC. For a non-separable Hamiltonian that is defined on a Riemannian manifold however, then Equation 3.77 and Equation 3.78 are defined *implicitly*, and we then need a further numerical scheme

to solve these steps. Many approaches exist, such as Newton’s method which makes use of derivative information, however such calculations are often costly in this context, particularly when the likelihood is expensive to compute. We therefore employ simple fixed point iterations run to convergence, and we find that just 5 or 6 iterations typically suffice for the statistical models we consider in Chapter 4.

The repeated application of the above steps provides the means to obtain a deterministic proposal that is guided not only by the derivative information of the target density, as in HMC and MALA, but also exploits the local geometric structure of the manifold as determined by the metric tensor. Intuitively, comparing the two Hamiltonians in Equations 2.87 and 3.74 shows that the constant mass matrix \mathbf{M} , defining a globally constant metric, is now replaced with the position specific metric. We now no longer need to tune the mass matrix \mathbf{M} , which can so dramatically affect the performance of HMC. Since the Generalised Leapfrog scheme is both time reversible and volume preserving, we can employ it as a proposal process to obtain a correct MCMC scheme that satisfies detailed balance (as shown in Chapter 2) and converges to the desired target density. The overall Riemannian Manifold HMC (RMHMC) scheme is given by Algorithm 3 and consists of drawing momentum values from a Gaussian distribution, whose covariance is the metric tensor given the current parameter values, and then running the Generalised Leapfrog integrator for a certain number of steps to give proposed moves $\boldsymbol{\theta}^*$ and \mathbf{p}^* , which are then accepted or rejected with probability $\min[1, \exp\{-H(\boldsymbol{\theta}^*, \mathbf{p}^*) + H(\boldsymbol{\theta}, \mathbf{p})\}]$. Just as for standard HMC, this sampling scheme produces an ergodic, time-reversible Markov chain satisfying detailed balance and whose stationary marginal density therefore gives us samples from $p(\boldsymbol{\theta})$. In this case however, there is no need to manually select and tune the mass matrix, since it is defined at each step by the underlying Riemannian geometry.

3.5 Population Manifold Methods

The manifold MCMC methods we have just introduced sample efficiently by exploiting the local geometry of the Riemannian manifold, however they will still potentially have difficulty sampling from multimodal distributions. This can be remedied by employing the manifold samplers within an overall parallel tempering or population Monte Carlo scheme (75, 96). The use of tempered distributions allows the Markov chains

to make more global steps between modes, and in addition the final samples may be used to accurately estimate the marginal likelihood for Bayesian model comparison via thermodynamic integration (27, 67, 118).

The tempered distributions are obtained by raising the likelihood to a power,

$$p(\boldsymbol{\theta}|\mathbf{y}, \beta) \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})^\beta \quad (3.80)$$

where $0 \leq \beta \leq 1$. If the metric tensor is given by the Expected Fisher Information then, given a uniform prior, the metric tensor for the tempered distribution has a particularly simple form,

$$G(\boldsymbol{\theta}, \beta) = \beta G(\boldsymbol{\theta}) \quad (3.81)$$

since the tempered log-likelihood is simply premultiplied by the power β . We shall employ this type of sampling scheme using manifold MCMC methods in Chapters 5 and 6 for the purpose of estimating marginal likelihoods for statistical models based on systems of ordinary differential equations. An example of tempered distributions generated from an ODE model is shown in Figure 3.5. A theoretical and computational investigation of the optimal annealing scheme for β is presented in (26, 27). Based on the findings of this previous analysis we employ $N = 30$ tempered distributions, where the values of $0 < \beta_n < 1$ are given by the formula

$$\beta_n = \left(\frac{n-1}{N-1} \right)^5 \quad (3.82)$$

We find that such an annealing scheme is robust to the choice of N for reasonably small $N > 20$ (27) and this allows us to perform model ranking of statistical model hypotheses by obtaining low variance estimates of the required marginal likelihoods.

3.6 Conclusions

In this chapter we have explored the idea that MCMC algorithms may be defined on different geometries. Existing dynamical MCMC methods based on Langevin diffusions and Hamiltonian systems are implicitly defined in a Euclidean space, whereby distance is measured in terms of changes in the parameter values alone. Moving towards a Riemannian geometry allows us to take changes in probability mass into account when

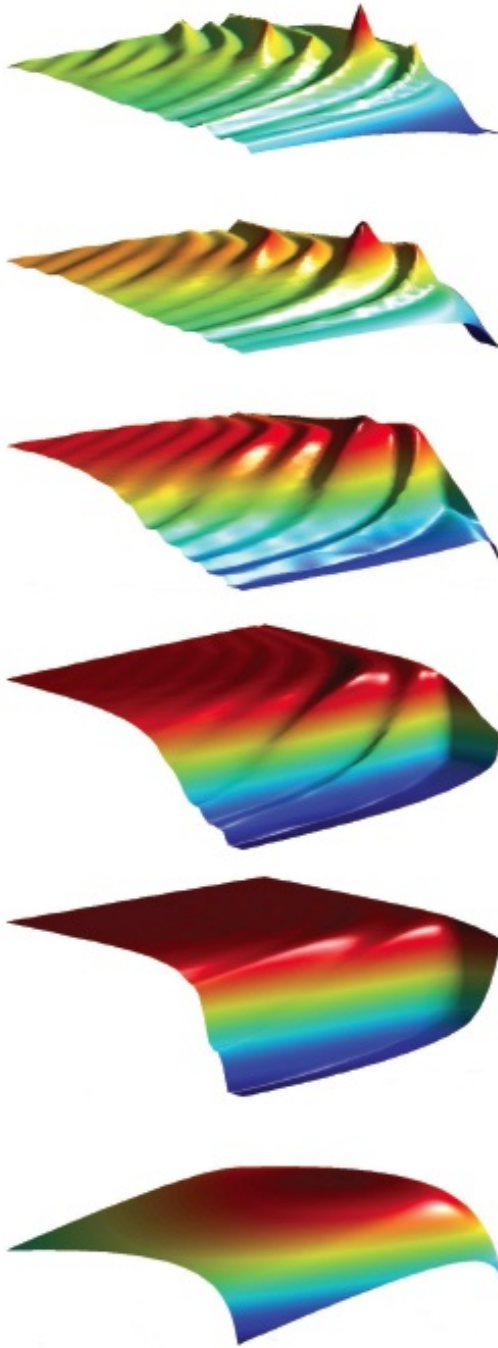


Figure 3.5: Example of tempered distributions - These 6 tempered distributions were obtained by employing the same ordinary differential equation model described in Figure 1.1 (26, 27). We observe that the highly nonlinear ripple structure in the posterior is gradually smoothed out until we reach the prior distribution. In population MCMC, a Markov chain is able to make moves between these distributions that satisfy detailed balance; if a chain becomes stuck in a local mode in the posterior, it may move down the distribution ladder until it reaches a smoother surface to move along, and in this way it may escape local modes and fully explore the parameter space.

defining distance based on a position specific metric tensor. The use of the Expected Fisher Information provides a natural link between statistical models, local sensitivity analysis and Riemannian geometry. In particular, we have developed generalisations of MALA and HMC methods based on a different geometry, which allows us to automatically propose moves that respect the local correlation structure induced by the statistical model. In the next chapter we shall evaluate and compare the performance of these new methods on a variety of models, and give some insight as to how they may be best employed. A summary of the manifold methods developed in this chapter is also given in Appendix A along with general guidelines for their use.

Algorithm 3 RMHMC with Generalised Leapfrog

```
1: Initialise current  $\theta$ 
2: for IterationNum = 1 to NumSamples do
3:   Sample new momentum  $\mathbf{p}^1$ 
4:   Calculate current  $H(\theta, \mathbf{p}^1)$ 
5:   Randomise  $N$  (leapfrog steps)
6:    $\theta^1 = \text{Current } \theta$ 
7:   for  $n = 1$  to  $N$  (leapfrog steps) do
8:     % Update the momentum with fixed point iterations
9:      $\hat{\mathbf{p}}^0 = \mathbf{p}^n$ 
10:    for  $i = 1$  to NumOfFixedPointSteps do
11:       $\hat{\mathbf{p}}^i = \mathbf{p}^n - \frac{\epsilon}{2} \nabla_{\theta} H(\theta^n, \hat{\mathbf{p}}^{i-1})$ 
12:    end for
13:     $\mathbf{p}^{n+\frac{1}{2}} = \hat{\mathbf{p}}^i$ 
14:    % Update the parameters with fixed point iterations
15:     $\hat{\theta}^0 = \theta^n$ 
16:    for  $i = 1$  to NumOfFixedPointSteps do
17:       $\hat{\theta}^i = \theta^n + \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\theta^n, \mathbf{p}^{n+\frac{1}{2}}) + \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\hat{\theta}^{i-1}, \mathbf{p}^{n+\frac{1}{2}})$ 
18:    end for
19:     $\theta^{n+1} = \hat{\theta}^i$ 
20:    % Update the momentum exactly
21:     $\mathbf{p}^{n+1} = \mathbf{p}^{n+\frac{1}{2}} - \frac{\epsilon}{2} \nabla_{\theta} H(\theta^{n+1}, \mathbf{p}^{n+\frac{1}{2}})$ 
22:  end for
23:  Calculate proposed  $H(\theta^N, \mathbf{p}^N)$ 
24:  LogRatio =  $-\log(\text{Proposed } H) + \log(\text{Current } H)$ 
25:  % Accept or reject according to Metropolis ratio
26:  if LogRatio > 0 or LogRatio > log(rand) then
27:    Set  $\theta = \theta^N$ 
28:  end if
29: end for
```

4

An Evaluation of Manifold MCMC Methods

We now investigate a number of applications of the differential geometric sampling methods that we developed in Chapter 3. We have already seen that RMHMC and mMALA methods appear to propose moves more effectively than HMC and MALA when sampling from a simple statistical model based on a 1-dimensional Gaussian distribution. We shall now examine in greater detail the efficiency of such methods on more challenging statistical models taking into account the computational costs of each approach. We note that the statistical models we look at in this chapter are purely for the purpose of comparing MCMC methodology; we are not so much interested in the particular interpretations of the statistical models themselves. Each chosen model has a unique set of characteristics that provide a challenge for any MCMC method and illuminate the advantages and disadvantages of each of the sampling approaches we employ. This chapter follows work that was published in (77).

4.1 Methods of Comparison

We employ a variety of MCMC methods to draw samples from the posterior distribution for each model. Each method is implemented in a similar fashion in the interpreted language Matlab to ensure as fair a comparison as possible, and we compare the relative efficiency of these sampling methods by calculating an effective sample size (ESS) using the posterior samples for each covariate. This measure gives the number of effectively

independent samples obtained from a larger collection of dependent samples and can be defined as,

$$ESS = \frac{N}{\left(1 + 2 \sum_{k=1}^K \gamma(k)\right)} \quad (4.1)$$

where N is the number of posterior samples and $\sum_k \gamma(k)$ is the sum of the K monotone sample autocorrelations as estimated by the initial monotone sequence estimator (74). Such an approach is also taken in (91), where the authors report the *mean* ESS, averaged over each of the covariates, however we feel this could give a rather inflated measure of the true ESS, since ideally we want a measure of the number of samples which are uncorrelated over *all* covariates. We therefore propose reporting the *minimum* ESS of the sampled covariates. Furthermore, we wish to take into account the computational complexity of each method, since it is plausible that a poorly performing method (in terms of ESS) may be very fast in obtaining samples and such a method could simply be run for a larger number of iterations to obtain better estimates. We therefore normalise the minimum ESS with respect to the CPU time and report the time taken to obtain 1 sample that is effectively uncorrelated across all covariates. Finally, the results reported for each of the models are the average values of 10 runs.

We shall now give a brief overview of the sampling methods that were used for comparison with the manifold-based sampling approaches.

4.1.1 Metropolis-Hastings

The simplest scheme implemented is an adaptive Metropolis-Hastings sampler; see for example Chapter 7 in (167), in which each covariate is updated individually with its stepsize being adapted every 100 iterations during the burn-in period to achieve an acceptance rate of between 20% and 50%. The stepsize is then fixed at the end of the burn-in period to ensure that samples are correctly drawn from the stationary distribution.

4.1.2 Metropolis Adjusted Langevin Algorithm

We implement a standard MALA sampler with proposed covariates being drawn from the multivariate normal distribution $\mathcal{N}(\boldsymbol{\theta}^* | \boldsymbol{\mu}(\boldsymbol{\theta}^n, \epsilon), \epsilon^2 \mathbf{I})$ as defined previously in Chapter 2. We follow the advice of (168) by scaling ϵ proportional to $D^{-\frac{1}{3}}$, where D is the

number of covariates, such that we obtain an acceptance rate of between 40% and 70%. We make use of any available information is available regarding the suitable choice of pre-conditioning matrix, as is the case for the log-Gaussian Cox example, otherwise we use MALA without pre-conditioning and focus on tuning the stepsize.

4.1.3 Hamiltonian Monte Carlo

In the following simulations we employ a fixed number of leapfrog steps and vary the stepsize manually for each dataset to achieve an acceptance rate of between 70% and 90%. This requires a number of exploratory runs of the algorithm, which we don't take into account when calculating computational time. As expected, the unit mass matrix we employ worked well for posterior distributions whose standard deviations for each covariate are of similar magnitudes, in other words when the parameter space is relatively isotropic as we saw in Chapter 2. We encode any information we have regarding the standard deviations of the marginal posterior distributions of the parameters in a diagonal mass matrix. More often than not, however, such information is not available and it is unclear how one might otherwise tune the mass matrix in any principled manner.

4.1.4 Manifold Methods

As we adopt a Bayesian approach in these examples, we wish to make proposals based on the local geometry of the posterior density. We therefore calculate the metric tensor employing the joint probability of data and parameters, such that

$$-E_{\mathbf{y}|\boldsymbol{\theta}} \left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log\{p(\mathbf{y}, \boldsymbol{\theta})\} \right] = -E_{\mathbf{y}|\boldsymbol{\theta}} \left[\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^2} \right] - \frac{\partial^2 \log(p(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2} \quad (4.2)$$

which is the Expected Fisher Information plus the negative Hessian of the log-prior. In this way we can capture the geometric structure of the prior in the metric tensor (63, 198).

4.2 Bayesian Logistic Regression

Logistic regression is a generalised linear model commonly used for binary regression (72, 125). It is a popular choice of model due to the regression coefficients having a

strong statistical interpretation in terms of changes in the log odds. Let us consider an $N \times D$ design matrix \mathbf{X} comprising N samples each with D covariates and a binary response variable $\mathbf{t} \in \{0, 1\}^N$. Denoting the logistic link function as $s(\cdot)$, a Bayesian logistic regression model is obtained by the introduction of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^D$ with an appropriate prior, which for illustrative purposes is given as $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$ where α is given. Neglecting constants, the log joint-likelihood follows in standard form as

$$\log p(\mathbf{t}, \boldsymbol{\beta} | \mathbf{X}, \alpha) = \mathcal{L}(\boldsymbol{\beta}) - \frac{1}{2\alpha} \boldsymbol{\beta}^\top \boldsymbol{\beta} = \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{t} - \sum_{n=1}^N \log(1 + \exp(\boldsymbol{\beta}^\top \mathbf{X}_{n,\cdot}^\top)) - \frac{1}{2\alpha} \boldsymbol{\beta}^\top \boldsymbol{\beta} \quad (4.3)$$

where $\mathbf{X}_{n,\cdot}$ denotes the vector that is the n^{th} row of the $N \times D$ matrix \mathbf{X} . The derivative of the log joint-likelihood is

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} - \alpha^{-1} \boldsymbol{\beta} \quad (4.4)$$

and its second derivative follows as

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}^2} - \alpha^{-1} \mathbf{I} \quad (4.5)$$

which comprises the matrix of second derivatives of the likelihood and the log-prior. This particular model induces a non-constant metric tensor, and a variety of datasets were used to provide a challenging test for any choice of sampling method; the number of covariates ranges from 2 to 24, and the number of data points ranges from 250 to 1000. We form the metric tensor based on the expected Fisher Information plus the negative Hessian of the log-prior in order to include the effect of the prior on the geometry, although in this particular model the expected and observed Fisher information are the same. The metric tensor therefore follows as

$$G(\boldsymbol{\beta}) = \mathbf{X}^\top \boldsymbol{\Lambda} \mathbf{X} + \alpha^{-1} \mathbf{I} \quad (4.6)$$

where the diagonal $N \times N$ matrix $\boldsymbol{\Lambda}$ has elements $\Lambda_{n,n} = s(\boldsymbol{\beta}^\top \mathbf{X}_{n,\cdot}^\top)(1 - s(\boldsymbol{\beta}^\top \mathbf{X}_{n,\cdot}^\top))$ where $\mathbf{X}_{n,\cdot}$ denotes the vector that is the n^{th} row of the $N \times D$ matrix \mathbf{X} . Finally the derivative matrices of the metric tensor take the form

Table 4.1: Summary of datasets for Bayesian logistic regression

Name	Covariates (D)	Data Points (N)	Dimension of β (b)
Pima Indian	7	532	8
Australian Credit	14	690	15
German Credit	24	1000	25
Heart	13	270	14
Ripley	2	250	7

$$\frac{\partial G(\beta)}{\partial \beta_i} = \mathbf{X}^\top \mathbf{\Lambda} \mathbf{V}^i \mathbf{X} \quad (4.7)$$

where the $N \times N$ diagonal matrix \mathbf{V}^i has elements $(1 - 2s(\beta^\top \mathbf{X}_{n,\cdot}^\top))X_{ni}$. The above identities are all that are required to define the RMHMC and mMALA sampling methods, which will be illustrated in the following experimental section.

4.2.1 Experimental Results for Bayesian Logistic Regression

We present results from the analysis of 5 datasets (135, 166), summarised in Table 4.1. Although the manifold based methods can easily cope with the raw data, we follow standard practice and normalise the datasets such that each covariate has zero mean and a standard deviation of one. This allows a fair comparison with other sampling methods, which would generally run into numerical problems with unnormalised data due to the non-isotropic nature of the resulting posterior distribution.

Given each dataset, we wish to sample from the posterior distribution over the regression coefficients β , and in each experiment vague Gaussian prior distributions were employed such that $\pi(\beta_i) \sim \mathcal{N}(0, 100)$. A linear logistic regression model with intercept was used for each of the datasets with the exception of the Ripley dataset, for which a cubic polynomial regression model was employed (166).

We reproduce the results of Holmes and Held (91) by allowing 5000 burn in iterations, such that each sampler reaches the stationary distribution and has time to adapt as necessary. The next 5000 iterations were used to collect posterior samples and the CPU time that each method took to collect these samples was recorded. We investigate the use of RMHMC, mMALA, HMC, MALA and Metropolis-Hastings applied to this

problem, and in addition we implement the following sampling methods that have been previously suggested in the literature for logistic regression models.

4.2.1.1 Auxiliary Variable Gibbs Sampler

The auxiliary variable Gibbs sampler of Holmes and Held (91) was implemented with a joint update of $\{\mathbf{z}, \boldsymbol{\beta}\}$, where $\mathbf{z} \in \mathbb{R}^N$ is the auxiliary variable designed to improve mixing of the covariate samples. We implemented the algorithm based on the very detailed pseudo-code given in the appendix of their paper, and in contrast to the M-H algorithm this method has the advantage of requiring no tuning of parameters. The main computational expense however is in the repeated sampling from truncated normal distributions, for which we implemented code based on the efficient method defined in (100).

4.2.1.2 Iterated Weighted Least Squares

We consider in addition the second order method Iterated Weighted Least Squares (IWLS) (68), which makes use of second derivatives in its Metropolis-Hastings proposal steps. It should be noted that IWLS is equivalent to simplified mMALA but without a tuneable stepsize. This method is relatively straightforward to implement and does not allow any tuning, similar to the auxiliary variable Gibbs sampler (91).

4.2.2 Comparison of MCMC Methods

We begin by investigating the RMHMC method in detail for the most challenging of our five datasets, German Credit, which consists of 24 covariates and 1000 datapoints. We then compare the results for all datasets employing the alternative sampling methods described previously.

Since RMHMC automatically adapts the full mass matrix via the metric tensor depending on its current position, we consider fixing ϵ and adjusting the number of leapfrog steps. Table 4.2 shows the results of the generalised leapfrog integration scheme using a variety of choices for these parameters. We found that sampling generally became more efficient as L , was increased, i.e. when the chain could traverse a greater distance. RMHMC allows for larger integration steps to be used and the value of 0.5

Table 4.2: RMHMC with generalised leapfrog integration scheme - investigating the effect of parameter settings on sampling efficiency with German Credit dataset

ϵL	Max ϵ	Mean Time (s)	Min ESS	s/Min ESS
1	0.5	131.7	674	0.195
2	0.5	193.6	2139	0.090
3	0.5	287.9	4791	0.060

was found to be a suitable choice for ϵ . The choice of 6 leapfrog steps was implemented for all datasets, resulting in an acceptance rate of between 70% and 90%.

We find that the RMHMC and mMALA sampling methods work very well for a variety of datasets and are fairly robust to the choice of algorithm parameters. For comparison with the alternative sampling methods, we chose the settings for RMHMC based on the above analysis. The scaling for mMALA was chosen to obtain an acceptance rate of between 50% and 70%. We repeated the sampling experiments 10 times and averaged the results, which are shown for each of the datasets in Tables 4.3 to 4.7.

All methods converge within 5000 burn-in iterations for all the normalised datasets. The manifold based methods generally outperform their equivalent, non-manifold counterparts, HMC and MALA, particularly for datasets that have stronger correlations between the covariates.

Figure 4.1 shows the trace and autocorrelation plots for 1000 posterior samples using the Heart dataset. The difference in autocorrelation is quite striking, both from inspection of the traces and from examination of the autocorrelation plots themselves. The autocorrelation of the RMHMC samples drop towards zero far quicker than any of the other methods.

As the number of covariates in the dataset increases, so the overall performance of RMHMC and mMALA decreases. This is due to the increased computational burden of calculating partial derivatives of the metric tensor with respect to each of the covariates. It is clear that for logistic regression problems with a very high number of covariates, for example in excess of 100 covariates, the use of RMHMC and mMALA will become computationally infeasible since we have to compute the partial derivatives of the metric tensor with respect to each of the covariates. As an alternative we can use a simplified mMALA scheme that assumes a locally constant metric tensor, avoiding the need to

Table 4.3: Comparison of sampling methods with Australian Credit dataset, $D = 14$, $N = 690$, 15 regression coefficients

Method	Time	ESS (Min, Med, Max)	s/Min ESS	Rel. Speed
Metropolis	10.8	(314, 709, 979)	0.034	$\times 320$
Aux. Var.	407.5	(37.4, 1054, 1405)	10.9	$\times 1$
MALA	2.7	(22.3, 576.8, 990.6)	0.122	$\times 89.3$
HMC	87.3	(3197, 3612, 3982)	0.027	$\times 403$
IWLS	5.15	(130, 215, 346)	0.040	$\times 272$
mMALA	11.7	(702, 867, 1037)	0.0167	$\times 652$
mMALA Simp.	3.2	(487, 625, 746)	0.006	$\times 1817$
RMHMC	81.7	(4975, 5000, 5000)	0.016	$\times 681$
RMHMC (Stud. t)	87.3	(1083, 1625, 2002)	0.081	$\times 134$
RMHMC (Fixed)	6.7	(4125, 4948, 5000)	0.0016	$\times 6812$

calculate derivatives. Similarly, we could also employ an RMHMC sampling scheme in which the metric tensor is fixed once the chain has converged to the stationary distribution; this is equivalent to standard HMC with a preconditioning matrix given by the Expected Fisher Information at some point in the posterior distribution.

We also consider an alternative second order method, IWLS, which makes use of terms involving second derivatives and therefore some measure of the curvature of the parameter space. IWLS performs fairly poorly however; indeed in the examples it performs about the same as parameter-wise Metropolis-Hastings. The flexibility of having an adjustable stepsize makes a great difference to the overall efficiency, which becomes apparent when we compare IWLS with the results of simplified mMALA method.

4.2.3 Comparison of mMALA and RMHMC Variants

We now investigate variants of RMHMC and mMALA to see whether results may be improved based on slight alterations to the standard forms. We first consider a simplified version of mMALA, which assumes a locally flat metric tensor during each Metropolis step and will still converge to the stationary distribution due to the Metropolis-Hastings adjustment. It is clear that this is computationally much less expensive than the full mMALA as it avoids the calculation of metric tensor derivatives. It is interesting that

4.2 Bayesian Logistic Regression

Table 4.4: Comparison of sampling methods with German Credit dataset, $D = 24$, $N = 1000$, 25 regression coefficients

Method	Time	ESS (Min, Med, Max)	s/Min ESS	Rel. Speed
Metropolis	23.4	(167, 613, 1015)	0.140	$\times 4.4$
Aux. Var.	618.8	(1006, 2211, 2640)	0.614	$\times 1$
MALA	3.5	(95.5, 316, 667)	0.037	$\times 16.6$
HMC	117.9	(3182, 3632, 3986)	0.037	$\times 16.6$
IWLS	9.3	(253, 572, 918)	0.037	$\times 16.6$
mMALA	42.3	(604, 766, 902)	0.070	$\times 8.8$
mMALA Simp.	5.0	(435, 615, 747)	0.012	$\times 51.2$
RMHMC	246.6	(4757, 5000, 5000)	0.052	$\times 11.8$
RMHMC (Stud. t)	257.4	(3981, 4934, 5000)	0.065	$\times 9.4$
RMHMC (Fixed)	8.6	(4409, 5000, 5000)	0.0019	$\times 323$

Table 4.5: Comparison of sampling methods with Pima Indian dataset, $D = 7$, $N = 532$, 8 regression coefficients

Method	Time	ESS (Min, Med, Max)	s/Min ESS	Rel. Speed
Metropolis	3.8	(347, 552, 980)	0.011	$\times 19.3$
Aux. Var.	304.3	(1432, 1888, 2295)	0.212	$\times 1$
MALA	1.56	(316, 550, 895)	0.005	$\times 42.4$
HMC	45.7	(3265, 3605, 3893)	0.014	$\times 15.1$
IWLS	2.6	(1386, 1937, 2379)	0.0019	$\times 111.6$
mMALA	4.2	(1135, 1286, 1412)	0.0037	$\times 57.3$
mMALA Simp.	1.9	(1046, 1160, 1300)	0.0018	$\times 117.8$
RMHMC	34.4	(5000, 5000, 5000)	0.0069	$\times 30.7$
RMHMC (Stud. t)	38.6	(3928, 4432, 4688)	0.0098	$\times 21.6$
RMHMC (Fixed)	5.8	(4912, 5000, 5000)	0.0012	$\times 177$

4.2 Bayesian Logistic Regression

Table 4.6: Comparison of sampling methods with Heart dataset, $D = 13$, $N = 270$, 14 regression coefficients

Method	Time	ESS (Min, Med, Max)	s/Min ESS	Rel. Speed
Metropolis	4.4	(418, 637, 905)	0.010	$\times 21.2$
Aux. Var.	150.9	(711, 1233, 1676)	0.212	$\times 1$
MALA	1.1	(279, 524, 814)	0.0038	$\times 55.8$
HMC	27.6	(3246, 3647, 4003)	0.0085	$\times 24.9$
IWLS	2.4	(87, 186, 381)	0.028	$\times 7.6$
mMALA	5.6	(656, 789, 903)	0.0085	$\times 24.9$
mMALA Simp.	1.6	(371, 481, 617)	0.0043	$\times 49.3$
RMHMC	42.2	(4862, 5000, 5000)	0.0087	$\times 24.4$
RMHMC (Stud. t)	48.0	(2603, 2904, 3171)	0.018	$\times 11.8$
RMHMC (Fixed)	3.4	(4403, 4940, 5000)	0.00076	$\times 279$

Table 4.7: Comparison of sampling methods with Ripley dataset, $D = 2$, $N = 250$, 7 regression coefficients

Method	Time	ESS (Min, Med, Max)	s/Min ESS	Rel. Speed
Metropolis	2.1	(59, 99, 271)	0.035	$\times 201$
Aux. Var.	139.6	(19, 44, 283)	7.06	$\times 1$
MALA	0.97	(33, 58, 101)	0.029	$\times 253$
HMC	24.8	(3326, 3719, 4053)	0.0076	$\times 967$
IWLS	1.46	(101, 207, 328)	0.015	$\times 490$
mMALA	3.3	(447, 579, 685)	0.0075	$\times 980$
mMALA Simp.	1.3	(291, 403, 473)	0.0045	$\times 1633$
RMHMC	28.0	(4273, 4677, 4961)	0.0065	$\times 1131$
RMHMC (Stud. t)	31.9	(2829, 3088, 3289)	0.011	$\times 668$
RMHMC (Fixed)	3.0	(1917, 3572, 4204)	0.0016	$\times 4413$

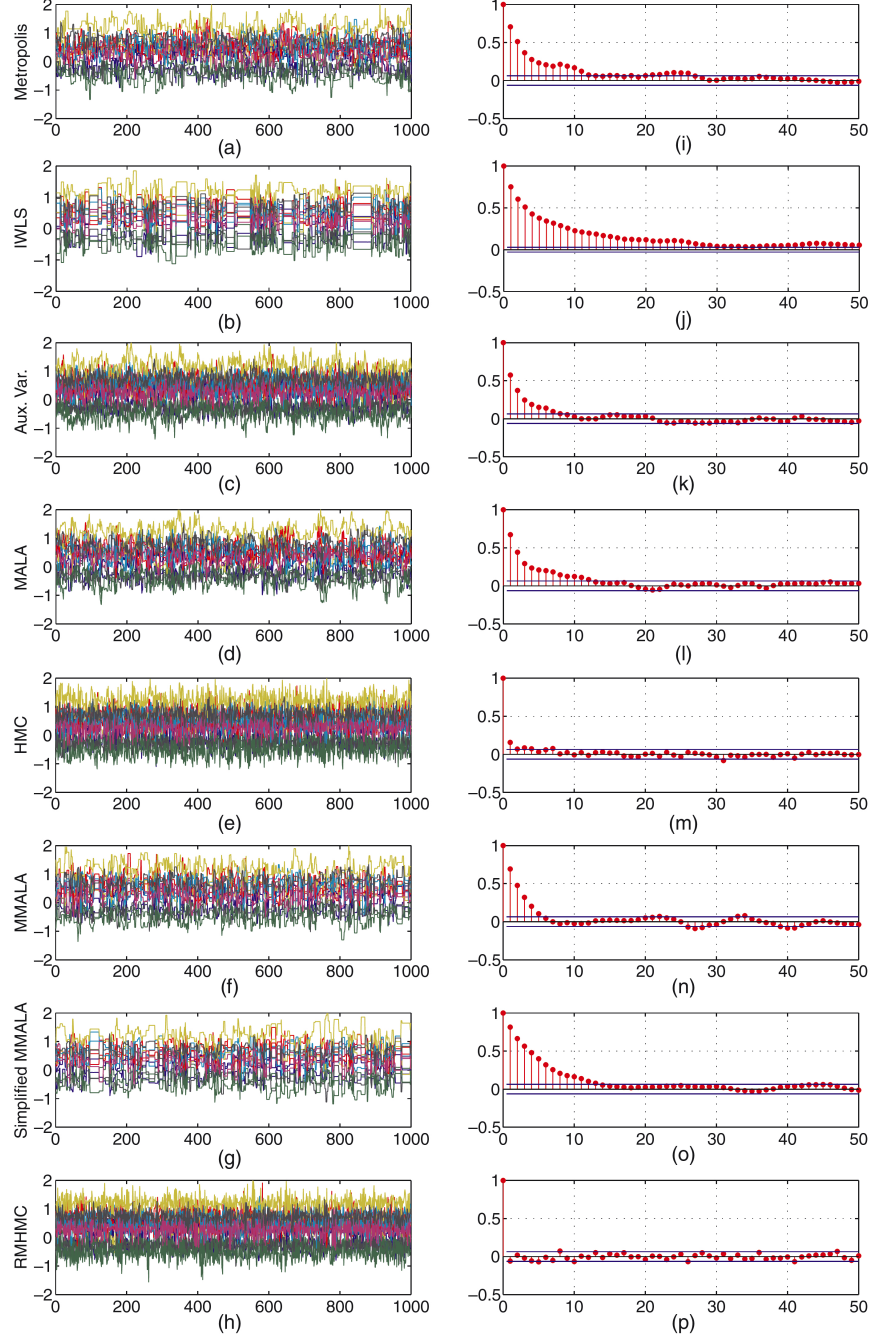


Figure 4.1: Autocorrelation of samples for Bayesian logistic regression with the Heart dataset - Trace plots for 1000 posterior samples with the Heart dataset using (from top to bottom) Metropolis, IWLS, auxiliary variable sampler, standard MALA, standard HMC, mMALA, Simplified mMALA and RMHMC. Autocorrelation plots are also shown for the first sampled covariate.

simplified mMALA has worse ESS than the full mMALA scheme, which intuitively makes sense since proposed steps across the manifold are approximated by not taking into account the rate of change of the local metric. The time normalised ESS however is far better due to the computational cost being much smaller. For this statistical model the best scheme in terms of the time normalised ESS is RMHMC with the metric tensor being fixed at the end of the burn-in period, since the geometry does not change much within the high probability region of the posterior distribution and it is mainly during the burn in phase that the metric tensor changes most rapidly.

It is also interesting to investigate the use of an alternative kinetic energy function in RMHMC. This idea is briefly mentioned in (125) although no example is given. We therefore consider the use of a kinetic energy term based on the Student- t density, with the idea that since the heavy tails might occasionally mean a larger momentum is sampled, this could plausibly result in lower correlation in the samples drawn from the target distribution. We note that since the multivariate Student- t distribution is symmetric, then the resulting Hamiltonian is still reversible. The simulations take slightly longer to run than with standard Gaussian distributed momentum using the same integration time steps. This is due to the increased computation required to sample from a Student- t distribution, and also the fact that there is more involved computation required to calculate the dynamics of this new Hamiltonian. The results show that the ESS is actually significantly less than that of a Hamiltonian defined with Gaussian momentum. This is possibly a result of a higher concentration of mass producing momenta with values closer to zero, even though there will be occasional samples of momentum with much larger magnitude.

In our simulation study manifold based methods perform extremely well with this statistical model compared to the other methods, when using small to medium sized datasets. It is interesting to note that due to the dense matrix form of the metric tensor and its inverse, the computational cost of mMALA and RMHMC on Bayesian logistic regression will not scale favourably; it can be seen that their time-normalised efficiency does indeed decrease as the number of regression coefficients in the dataset increases. This issue of scaling can however be eased somewhat by employing simplified mMALA sampling, which assumes a locally constant metric tensor and thus avoids the expensive computation of the metric tensor derivatives, or by employing RMHMC with a fixed mass matrix after the burn-in period.

4.3 Stochastic Volatility Model

We now consider the stochastic volatility model (SVM) studied in (106, 125) which takes the form of a latent variable model. It is defined with the latent volatilities taking the form of a 1st order autoregressive process (AR(1) process). The observations are defined as

$$y_t = \epsilon_t \beta \exp(x_t/2) \quad (4.8)$$

and the latent volatilities take the form

$$x_{t+1} = \phi x_t + \eta_{t+1} \quad (4.9)$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$, $\eta_t \sim \mathcal{N}(0, \sigma^2)$ and $x_1 \sim \mathcal{N}(0, \sigma^2/(1 - \phi^2))$. The joint probability therefore follows as

$$p(\mathbf{y}, \mathbf{x}, \beta, \phi, \sigma) = \prod_{t=1}^T p(y_t|x_t, \beta) p(x_1) \prod_{t=2}^T p(x_t|x_{t-1}, \phi, \sigma) \pi(\beta) \pi(\phi) \pi(\sigma) \quad (4.10)$$

We split up the sampling procedure into two steps, which allows the implementation of both mMALA and RMHMC in a computationally efficient manner. Firstly we simulate ϕ, σ, β from $p(\beta, \phi, \sigma|\mathbf{y}, \mathbf{x})$, where the priors are chosen to be $p(\beta) \propto 1/\beta$, $\sigma^2 \sim \text{Inv-}\chi^2(10, 0.05)$ and $(\phi + 1)/2 \sim \text{Beta}(20, 1.5)$. The variable ϕ takes values between -1 and 1 , and σ must be a positive real number. One way to deal with such constraints is to implement transformations of these to the real line, which we describe in the next subsection, noting that this introduces a Jacobian factor into the acceptance ratio in the standard manner when transforming probability distributions. Secondly we sample the latent volatilities by simulating from the conditional $p(\mathbf{x}|\mathbf{y}, \beta, \sigma, \phi)$. We shall consider the use of mMALA, RMHMC, MALA and HMC for the purpose of sampling both the parameters and latent volatilities.

4.3.1 mMALA and RMHMC for SVM Parameters

We require the partial derivatives of the joint log probability with respect to the transformed parameters to implement MALA and HMC, as well as expressions for the metric tensor and its partial derivatives, in order to employ mMALA and RMHMC. All of these quantities may be obtained straightforwardly. We then use these methods to draw samples from the conditional posterior $p(\beta, \alpha, \gamma | \mathbf{y}, \mathbf{x},)$.

We employ the transformations $\sigma = \exp(\gamma)$ and $\phi = \tanh(\alpha)$ to deal with constrained parameters. The derivatives of the transformations follow as $\frac{d\sigma}{d\gamma} = \exp(\gamma) = \sigma$ and $\frac{d\phi}{d\alpha} = 1 - \tanh^2(\alpha) = 1 - \phi^2$. The partial derivatives of the joint log probability, $L = \log p(\mathbf{y}, \mathbf{x} | \beta, \sigma, \phi)$, with respect to the transformed parameters are as follows

$$\frac{\partial L}{\partial \beta} = -\frac{T}{\beta} + \sum_{t=1}^T \frac{y_t^2}{\beta^3 \exp(x_t)} \quad (4.11)$$

$$\frac{\partial L}{\partial \gamma} = \frac{\partial L}{\partial \sigma} \frac{d\sigma}{d\gamma} = -T + \frac{x_1^2(1 - \phi^2)}{\sigma^2} + \sum_{t=2}^T \frac{(x_t - \phi x_{t-1})^2}{\sigma^2} \quad (4.12)$$

$$\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial \phi} \frac{d\phi}{d\alpha} = -\phi + \frac{\phi x_1^2(1 - \phi^2)}{\sigma^2} + \sum_{t=2}^T \frac{x_{t-1}(x_t - \phi x_{t-1})(1 - \phi^2)}{\sigma^2} \quad (4.13)$$

If we want to sample the parameters using mMALA or RMHMC, then we also need expressions for the metric tensor and its partial derivatives with respect to β, σ and ϕ . We can obtain the following expressions for the individual components of the metric tensor for the log probability density

$$\begin{aligned} E \left\{ \frac{\partial^2 L}{\partial \beta^2} \right\} &= -\frac{2T}{\beta^2}, \quad E \left\{ \frac{\partial^2 L}{\partial \gamma^2} \right\} = -2T, \quad E \left\{ \frac{\partial^2 L}{\partial \beta \partial \gamma} \right\} = E \left\{ \frac{\partial^2 L}{\partial \beta \partial \alpha} \right\} = 0 \\ E \left\{ \frac{\partial^2 L}{\partial \gamma \partial \alpha} \right\} &= -2\phi, \quad E \left\{ \frac{\partial^2 L}{\partial \alpha^2} \right\} = -2\phi^2 - (T-1)(1 - \phi^2) \end{aligned}$$

Thus the expected Fisher information and its partial derivatives follow as

$$\begin{aligned} G(\alpha, \gamma, \beta) &= \begin{bmatrix} \frac{2T}{\beta^2} & 0 & 0 \\ 0 & 2T & 2\phi \\ 0 & 2\phi & 2\phi^2 + (T-1)(1 - \phi^2) \end{bmatrix}, \quad \frac{\partial \mathbf{G}}{\partial \beta} = \begin{bmatrix} -\frac{4T}{\beta^3} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ \frac{\partial \mathbf{G}}{\partial \gamma} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \frac{\partial \mathbf{G}}{\partial \alpha} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2\phi(3 - T)(1 - \phi^2) \end{bmatrix} \end{aligned}$$

We therefore require expressions for the second order derivatives of the log priors to get the overall metric tensor, and also the third order derivatives of the log priors to calculate the partial derivatives of the metric tensor, which follow straightforwardly.

4.3.2 mMALA and RMHMC for SVM Latent Volatilities

Let us now consider the required expressions for sampling the latent volatilities. Defining $\mathbf{u} = (x_3, \dots, x_T)^\top$, $\mathbf{v} = (x_2, \dots, x_{T-1})^\top$, $\mathbf{w} = \frac{\phi}{\sigma^2}(\mathbf{u} - \phi\mathbf{v})$, $\mathbf{s} = (s_1, \dots, s_T)^\top$ such that $s_i = 0.5(1 - y_i^2\beta^{-2}\exp(-x_i))$, $\delta_1 = -\sigma^{-2}(x_1 - \phi x_2)$, and $\delta_T = -\sigma^{-2}(x_T - \phi x_{T-1})$, we define the vector $\mathbf{r} = (\delta_1, \mathbf{w}^\top, \delta_T)^\top$ and the gradient $\nabla_{\mathbf{x}} \log p(\mathbf{y}, \mathbf{x} | \beta, \phi, \sigma) = \mathbf{s} - \mathbf{r}$.

To devise mMALA and RMHMC samplers for the latent volatilities, \mathbf{x} , we also require an expression for the metric tensor and its partial derivatives with respect to the latent volatilities. For the data probability of the model, $p(\mathbf{y} | \mathbf{x}, \beta)$, the expected Fisher Information is the scaled identity matrix $\frac{1}{2} \times \mathbf{I}$. The latent volatility is an AR(1) process having covariance matrix \mathbf{C} with elements $E\{x_{t+n}x_t\} = \phi^{|n|}\sigma^2/(1 - \phi^2)$ (173) and as in the previous examples the metric tensor is defined as the sum of the expected Fisher Information and the negative Hessian of the log-prior, $\mathbf{G} = \frac{1}{2} \times \mathbf{I} + \mathbf{C}^{-1}$, conditional on current values of σ, ϕ, β . Now the expression for the covariance matrix is completely dense and is therefore computationally expensive to manipulate. Fortunately, this AR(1) process admits a simple analytic expression for the precision matrix in the form of a sparse tridiagonal matrix, such that the diagonal elements are equal to $(1 + \phi^2)/\sigma^2$, with the exception of the first and last diagonal elements which are equal to $1/\sigma^2$, and the super and sub diagonal elements are equal to $-\phi/\sigma^2$. Thus the metric tensor also has a tridiagonal form. For large numbers of observations this sparse structure allows great gains in computational efficiency, since the inverse of this tridiagonal metric tensor may be computed in $\mathcal{O}(T \log T)$ as opposed to the usual $\mathcal{O}(T^3)$. We note that computationally efficient methods for manipulating tridiagonal matrices are automatically implemented by the standard routines in Matlab.

We notice that the metric tensor in this case is not a function of the latent volatilities \mathbf{x} and so the associated partial derivatives with respect to the latent volatilities are zero. In this case, the manifold is a simple inner product space and the RMHMC scheme is effectively HMC where the mass matrix \mathbf{M} is now defined based on the natural Riemannian geometry using the globally constant metric tensor \mathbf{G} . Likewise mMALA collapses to a MALA scheme preconditioned by the constant matrix \mathbf{G}^{-1} . It is clear

4.3 Stochastic Volatility Model

Table 4.8: Comparison of sampling the parameters β , σ and ϕ after 20,000 posterior samples averaged over 10 runs with 2000 simulated observations, $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$

Method	Mean Time	ESS (β, σ, ϕ)	S.E. (β, σ, ϕ)	s/(Min ESS)	Rel. Speed
MALA	44.0	(19.1, 11.3, 30.1)	(1.9, 0.8, 2.1)	3.89	$\times 36.7$
HMC	424.8	(117, 81, 198)	(9.3, 3.1, 10.3)	5.19	$\times 27.5$
mMALA	2455.9	(17.2, 17.4, 44.5)	(2.8, 2.4, 9.2)	142.8	$\times 1$
RMHMC	329.4	(325, 139, 344)	(19.0, 7.3, 25.2)	2.37	$\times 60.3$

Table 4.9: Comparison of sampling the latent volatilities after 20,000 posterior samples averaged over 10 runs with 2000 simulated observations, $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$

Method	Mean Time	ESS (min, median, max)	s/(Min ESS)	Rel. Speed
MALA	44.0	(9.7, 16.7, 28.4)	4.53	$\times 7.5$
HMC	424.8	(409, 624, 1239)	1.04	$\times 32.9$
mMALA	2455.9	(71.8, 131.0, 329.8)	34.2	$\times 1$
RMHMC	329.4	(977, 1689, 3376)	0.34	$\times 100.6$

that this preconditioning will improve both the mixing and overall ESS, see e.g. (116) for a recent application of this type of preconditioning in MALA. Just as for RMHMC, the preconditioning matrix emerges naturally from the underlying geometric principles.

4.3.3 Experimental Results for Stochastic Volatility Model

We now compare the computational efficiency of RMHMC, mMALA, HMC and MALA for sampling both the parameters and the latent variables of the stochastic volatility model as previously defined, and the results are shown in Tables (4.8) and (4.9). 2000 observations were simulated from the model with the parameter values $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$ as given in (125). Using this data, 20,000 posterior samples were collected after a burn-in period of 10,000 samples. This sampling procedure was repeated 10 times. The efficiency of sampling the parameters and the latent volatilities was once again compared in terms of time normalised ESS. MALA was tuned such that the acceptance ratio was between 40% and 70%, and we note that in this example it was

necessary to use a different tuning for the transient phase, at the start of the simulation away from the mode, than for the stationary phase. HMC was implemented using 100 leapfrog steps and tuning the stepsize to obtain an acceptance rate of between 70% and 90%, which resulted in a stepsize of 0.015 for hyperparameters and a stepsize of 0.03 for the latent volatilities. RMHMC was implemented using a stepsize of 0.5 and 6 integration steps per parameter proposal, and a stepsize of 0.1 and 50 integration steps per volatility proposal, both of which resulted in an acceptance rate of between 70% and 90%.

In terms of sampling the hyperparameters, manifold methods offer little advantage over standard sampling approaches due to the small dimensionality of the problem and lack of strong correlation present. RMHMC and MALA give the best performance in terms of time normalised ESS. MALA exhibits a very poor ESS, however the computation time is also extremely small compared to the other two methods. RMHMC has the highest raw ESS, but has much more computational overhead compared to MALA. The posterior marginal densities of the hyperparameters are shown in Figure 4.2. When we consider sampling the latent variables, RMHMC offers greater advantages. In particular, it samples more efficiently than HMC, partly because of the computationally efficient tridiagonal structure of the metric tensor and partly because RMHMC follows the natural gradient through the parameter space and requires significantly fewer leapfrog iterations to explore the target density. An illustration of the contrast between HMC and RMHMC sampling of the parameters of this model is given in Figures 4.3 and 4.4. In this example, mMALA performs very badly due to the need to take a Cholesky decomposition of the inverse metric tensor of the latent variables, which is a dense matrix, compared to RMHMC which only requires use of the tridiagonal metric tensor. It should be noted that RMHMC again requires very little tuning compared to the other methods; unlike MALA it does not require different tuning in different parts of the parameter space, and unlike HMC it requires no manual setting of a mass matrix.

4.4 Log Gaussian Cox Model

As a final example in this chapter we examine the use of differential geometric methodology for inference in a log-Gaussian Cox point process model as detailed in (38). This

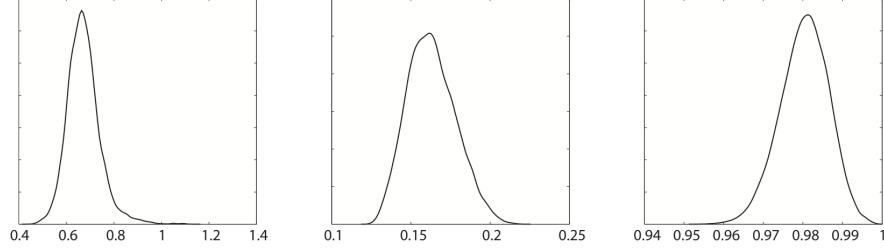


Figure 4.2: Posterior marginal distributions of the stochastic volatility model hyperparameters - Kernel smoothed posterior marginal densities for β , σ and ϕ respectively, employing RMHMC to draw 20,000 samples of the parameters and latent volatilities using a simulated dataset consisting of 2000 observations. The true values are $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$.

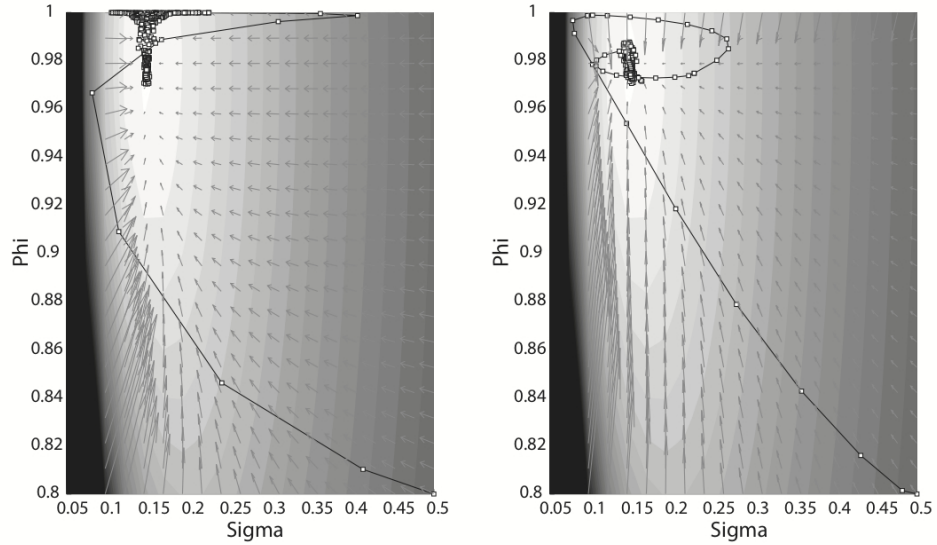


Figure 4.3: Comparison of sampling the latent volatilities in a stochastic volatility model using HMC and RMHMC - Contours plotted from the stochastic volatility model, in which the latent volatilities and the parameter β are set to their true values and the log-joint probability given different values of σ and ϕ is shown. The left hand plot shows the evolution of a Markov chain using HMC sampling with a unit mass matrix, the right hand plot shows the evolution of a chain from the same starting point using RMHMC sampling. We note that the use of the metric allows the RMHMC algorithm to converge much more quickly to the target density.

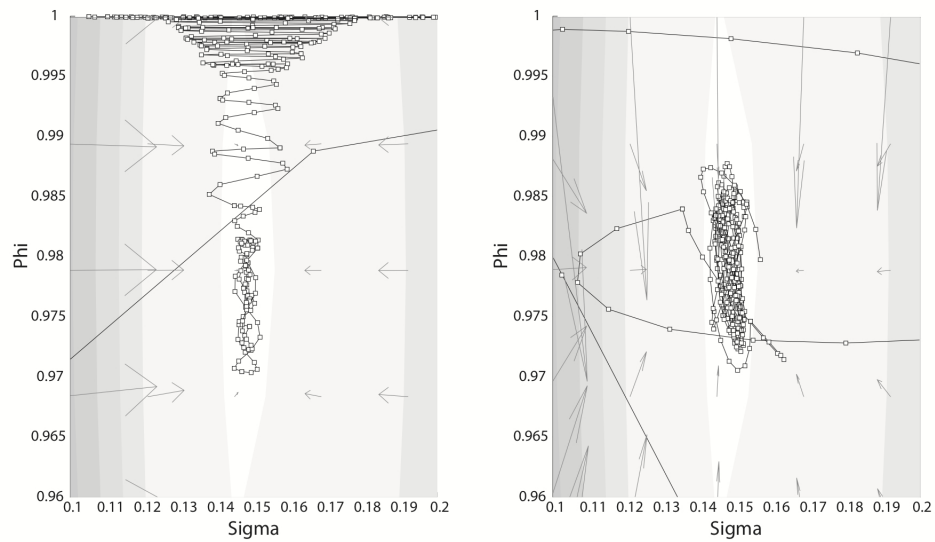


Figure 4.4: Close up of sampling the hyperparameters in a stochastic volatility model using HMC and RMHMC - A close up is shown of the Markov chain paths from Figure 4.3. The RMHMC sampling effectively normalises the gradients in each direction, whereas HMC sampling, with a unit mass matrix, exhibits stronger gradients along the horizontal direction compared to the vertical direction and therefore takes longer to converge to the target density. A carefully tuned mass matrix may improve HMC sampling, whereas RMHMC does this automatically by taking into account the local geometry.

is a particularly insightful example of another latent variable model where the target density is of high dimension with strong correlations, which provides a severe test of MCMC capability. The data, model and experimental protocol as described in (38) is adopted here. A 64×64 grid is overlaid on the area $[0, 1]^2$ with the number of points in each grid cell denoted by the random variables $\mathbf{Y} = \{Y_{i,j}\}$ which are assumed conditionally independent, given a latent intensity process $\Lambda(\cdot) = \{\Lambda(i, j)\}$, and are Poisson distributed with means $m\Lambda(i, j) = m \exp(X_{i,j})$, where $m = 1/4096$, $\mathbf{X} = \{X_{i,j}\}$, $\mathbf{x} = \text{Vec}(\mathbf{X})$, and $\mathbf{y} = \text{Vec}(\mathbf{Y})$, with \mathbf{X} a Gaussian process having mean $E\{\mathbf{x}\} = \mu \mathbf{1}$, and covariance function $\Sigma_{(i,j),(i',j')} = \sigma^2 \exp(-\delta(i, i', j, j')/64\beta)$, where $\delta(i, i', j, j') = \sqrt{(i - i')^2 + (j - j')^2}$. The joint density is

$$p(\mathbf{y}, \mathbf{x} | \mu, \sigma, \beta) \propto \prod_{i,j} \exp\{y_{i,j} x_{i,j} - m \exp(x_{i,j})\} \exp(-(\mathbf{x} - \mu \mathbf{1})^T \Sigma^{-1} (\mathbf{x} - \mu \mathbf{1})/2) \quad (4.14)$$

As in the previous example we consider an overall Gibbs scheme in which we alternately sample from $p(\mathbf{x} | \mathbf{y}, \sigma, \beta, \mu)$ and $p(\sigma, \beta | \mathbf{y}, \mathbf{x}, \mu)$. Denoting $\mathcal{L} \equiv \log p(\mathbf{y}, \mathbf{x} | \mu, \sigma, \beta)$ and $\mathbf{e} = \{m \exp(x_{i,j})\}$, then the derivative with respect to the latent variables follows straightforwardly as $\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{y} - \mathbf{e} - \Sigma^{-1} (\mathbf{x} - \mu \mathbf{1})$. The metric tensor follows as $-E_{\mathbf{y}, \mathbf{x} | \boldsymbol{\theta}} \{\nabla_{\mathbf{x}} \nabla_{\mathbf{x}} \mathcal{L}\} = \mathbf{\Lambda} + \Sigma^{-1}$, where the diagonal matrix $\mathbf{\Lambda}$, whose i th diagonal element is defined as $m \exp(\mu + (\Sigma)_{ii})$, follows from the expectation of the exponential of normal random variables. The metric tensor describing the manifold for the random field is constant $\mathbf{G} = \mathbf{\Lambda} + \Sigma^{-1}$. The mMALA and RMHMC schemes for the conditional, $p(\mathbf{x} | \mathbf{y}, \sigma, \beta, \mu)$, are then equivalent to MALA and HMC with mass and preconditioning matrices $\mathbf{M} = \mathbf{\Lambda} + \Sigma^{-1}$ and \mathbf{M}^{-1} , respectively. The computational cost of calculating the required inverse of the metric tensor scales as $\mathcal{O}(N^3)$, however once this quantity has been calculated, as for HMC, a large number of leapfrog steps may be made with little additional overhead, resulting in efficient sampling of the latent variables.

For sampling from the conditional $p(\sigma, \beta | \mathbf{y}, \mathbf{x}, \mu)$ we employ a metric tensor based on the expected Fisher information for the parameters $\boldsymbol{\theta} = [\sigma, \beta]$ which follows as $\mathbf{D}_{\boldsymbol{\theta}}$ whose (l, m) th element is $\frac{1}{2} \text{Tr}(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_l} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_m})$.

We employ a change of variables $\sigma^2 = \exp(\varphi_1)$ and $\beta = \exp(\varphi_2)$ to allow constrained sampling such that σ^2 and β are both strictly positive. The log-probability and gradients required for sampling the hyperparameters of the Gaussian process follow in standard form,

$$\frac{\partial \mathcal{L}}{\partial \varphi_i} = -\frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \right) + \frac{1}{2} (\mathbf{x} - \mu \mathbf{1})^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} (\mathbf{x} - \mu \mathbf{1}) \quad (4.15)$$

where $i = 1, 2$. The Fisher Information matrix also follows in standard form as

$$G(\varphi)_{ij} = \frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_j} \right) \quad (4.16)$$

Application of the standard expression for the derivative of trace operators provides an analytic expression for the derivative of the metric tensor with respect to the transformed parameters

$$\begin{aligned} \frac{\partial G(\varphi)_{ij}}{\partial \varphi_k} &= \frac{\partial}{\partial \varphi_k} \left[\frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_j} \right) \right] \\ &= -\frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_k} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_j} \right) + \frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \varphi_i \partial \varphi_k} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_j} \right) \\ &\quad - \frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_k} \Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_j} \right) + \frac{1}{2} \text{Tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \varphi_i} \Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \varphi_j \partial \varphi_k} \right) \end{aligned}$$

In our experiments we employ an infinitely differentiable stationary covariance function to calculate the $(i, j)^{th}$ entry of the covariance matrix,

$$\Sigma_{(i,j),(i',j')} = \sigma^2 \exp \left(-\frac{1}{64\beta} \delta(i, i', j, j') \right) \quad (4.17)$$

where $\delta(i, i', j, j') = \sqrt{(i - i')^2 + (j - j')^2}$. The gradients and the Fisher Information matrix above may therefore be obtained using the first and second partial derivatives of the covariance function. The first partial derivatives follow as,

$$\begin{aligned} \frac{\partial \Sigma_{i,j}}{\partial \varphi_1} &= \sigma^2 \exp \left(-\frac{1}{64\beta} \delta(i, i', j, j') \right) \\ \frac{\partial \Sigma_{i,j}}{\partial \varphi_2} &= \frac{\sigma^2}{64\beta} \exp \left(-\frac{1}{64\beta} \delta(i, i', j, j') \right) \delta(i, i', j, j') \end{aligned}$$

The second partial derivatives may also be easily calculated,

$$\begin{aligned}
\frac{\partial^2 \Sigma_{i,j}}{\partial \varphi_1^2} &= \sigma^2 \exp\left(-\frac{1}{64\beta} \delta(i, i', j, j')\right) \\
\frac{\partial^2 \Sigma_{i,j}}{\partial \varphi_1 \partial \varphi_2} &= \frac{\sigma^2}{64\beta} \exp\left(-\frac{1}{64\beta} \delta(i, i', j, j')\right) \delta(i, i', j, j') \\
\frac{\partial^2 \Sigma_{i,j}}{\partial \varphi_2^2} &= \frac{\sigma^2}{(64\beta)^2} \exp\left(-\frac{1}{64\beta} \delta(i, i', j, j')\right) \delta(i, i', j, j')^2 \\
&\quad - \frac{\sigma^2}{64\beta} \exp\left(-\frac{1}{64\beta} \delta(i, i', j, j')\right) \delta(i, i', j, j')
\end{aligned}$$

Once again we require expressions for the second order derivatives of the log priors to get the metric tensor over the full target distribution, and also the third order derivatives of the log priors to calculate the partial derivatives of the metric tensor. These follow straightforwardly from the $\text{Ga}(2, 0.5)$ priors employed over the hyperparameters σ^2 and β .

Noting that the metric tensor for the latent variables has dimension $N \times N$, where $N = 4096$ the $\mathcal{O}(N^3)$ operations required in the mMALA and RMHMC schemes are clearly going to be computationally costly. However, it should also be noted that in previous studies of this Log-Gaussian Cox process, (38), a transformation of the latent Gaussian field based on the Cholesky decomposition of $\Sigma^{-1} + \text{diag}(\mathbf{x})$ was also used, which therefore also scales as $\mathcal{O}(N^3)$.

It is possible to consider jointly sampling the hyperparameters and the latent variables. With $\mathcal{L} \equiv \log p(\mathbf{y}, \mathbf{x}, \sigma, \beta | \mu)$, we see that the Expected Fisher Information matrix is block diagonal with blocks $\mathbf{\Lambda} + \Sigma^{-1}$ and \mathbf{D}_{θ}^{-1} . Unfortunately, jointly sampling the latent variables and the hyperparameters proves to be computationally too costly to implement, as the metric tensor is now no longer fixed and so the generalised leapfrog integration scheme would have to be implemented in RMHMC with fixed point iterations, during each of which the metric tensor and its inverse have to be recalculated.

4.4.1 Experimental Results for Log-Gaussian Cox Processes

Following the example given in (38), we fix the parameters $\beta = 1/33$, $\sigma^2 = 1.91$ and $\mu = \log(126) - \sigma^2/2$. We generate a latent Gaussian field \mathbf{x} from the Gaussian process and use these values to generate count data \mathbf{y} from the latent intensity process $\mathbf{\Lambda}$. Given the generated data and the fixed hyperparameters, we infer \mathbf{x} using mMALA,

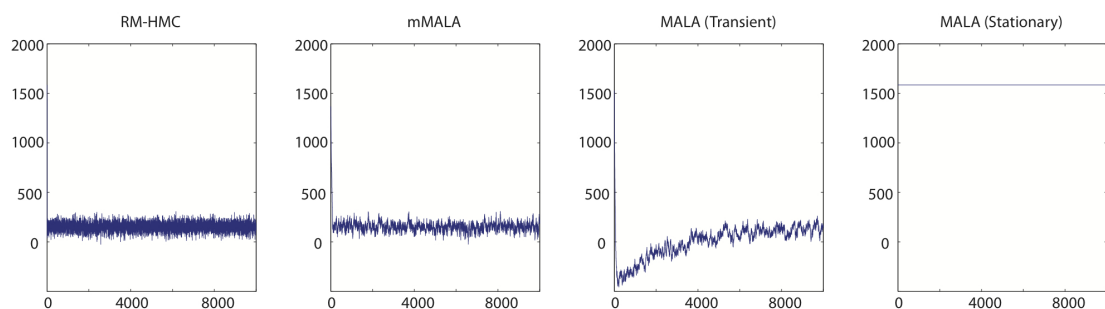


Figure 4.5: Trace plots of the log joint probability of the log-Gaussian Cox model - Trace plots of the log joint-probability for the first 10,000 samples of the latent variables of a log-Gaussian Cox process. The left hand plot shows the convergence of the RMHMC scheme which is able to directly sample the latent variables \mathbf{x} without the need for ad hoc reparameterisations and pilot runs for fine-tuning. The left-middle plot shows the convergence of the mMALA scheme which, since it also uses information about the manifold in the form of the metric tensor, is able to directly sample without any reparameterisations. The right-middle plot shows the log joint-probability for samples drawn by MALA using a reparameterisation of the latent variables. The scaling was carefully tuned to allow traversal of the parameter space to the posterior mode. The right hand plot shows the trace of the MALA sampler tuned for optimally sampling from the posterior mode. We note that the algorithm is now unable to traverse the parameter space when initialised away from this mode. Such fine-tuning and reparameterisation is frequently necessary when employing MALA.

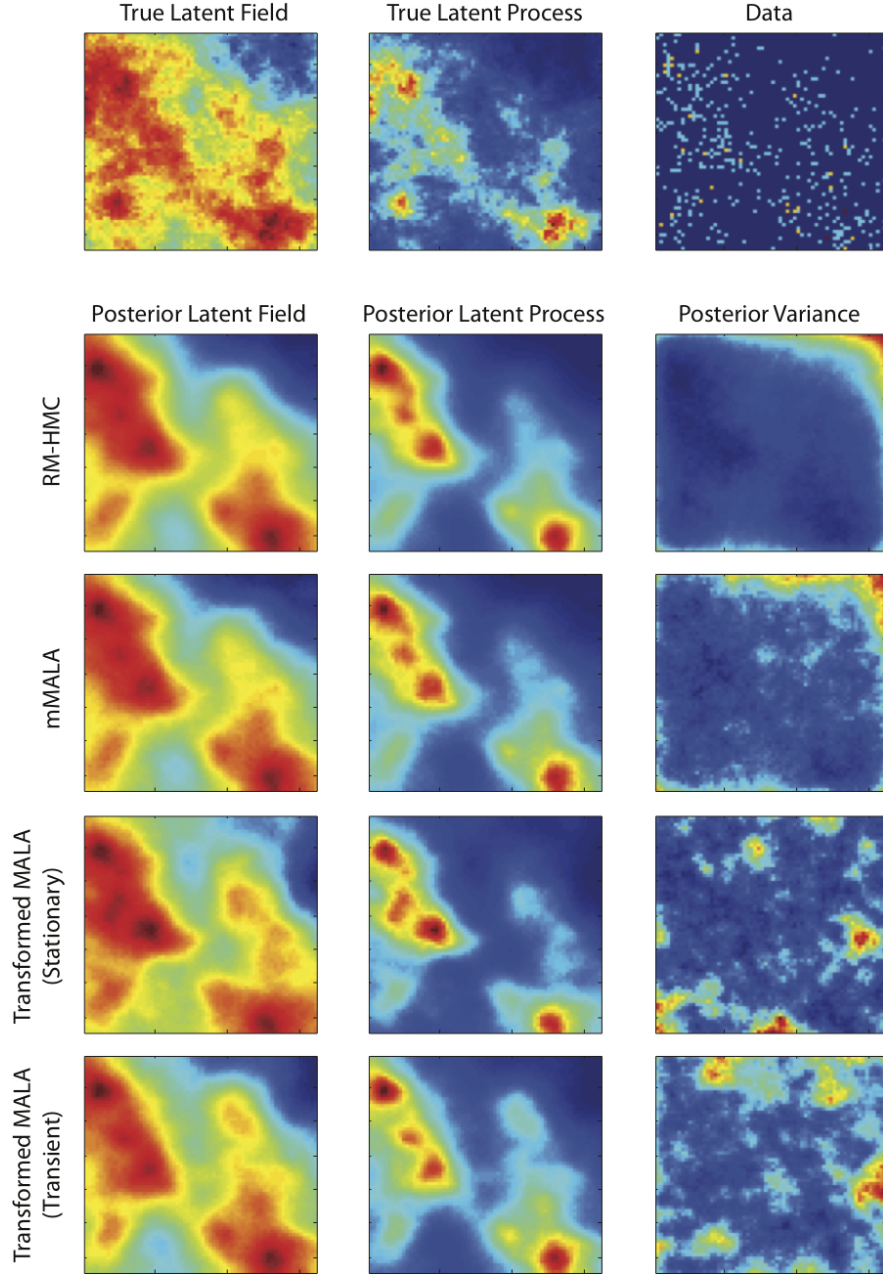


Figure 4.6: Posterior latent fields of the log-Gaussian Cox model - Posterior latent fields and processes and associated variance, using each of the sampling methods, compared to the true latent field and process. The data employed to infer the latent field are also shown in the top right plot. RMHMC produces the lowest variance estimates, which corresponds with it having the highest ESS. For RMHMC there is higher variance where there is less data, however for the other methods there are patchy areas of high variance due to correlations in the collected samples.

RMHMC and the MALA method as in (38). The algorithms were run on a single AMD Opteron processor with 8GB of memory and were again coded in Matlab for fairness of comparison.

In many settings MALA, like HMC, is particularly sensitive to the choice of scaling and very often a reparameterisation of the target density is required for these methods to be effective. Indeed this is seen to be the case with this particular example, where MALA is unable to sample \mathbf{x} directly. We therefore follow (38) and employ the transformation $\mathbf{X} = \mu\mathbf{1} + \mathbf{L}\mathbf{\Gamma}$, where \mathbf{L} is obtained by Cholesky factorisation such that $\{\Sigma + \text{diag}(\mathbf{x})\}^{-1} = \mathbf{L}\mathbf{L}^T$. Even after this reparameterisation, it is still necessary to carefully tune the scaling factor for this method to work at all. This challenging aspect of employing MALA has been investigated in detail in (38) who characterise the problem very well, advising great care in its implementation, but are ultimately unable to offer any panacea. In contrast to the necessary transformation and fine-tuning required by MALA, both mMALA and RMHMC allow us to directly sample the latent variables \mathbf{x} *without* having to manually choose a reparameterisation of the target density.

Figure 4.5 shows the traces of the log joint-probability for both methods using the starting position $x_{i,j} = \mu$ for $i, j = 1, \dots, 64$. Note that for MALA these starting positions must be transformed into corresponding values for $\mathbf{\Gamma}$. The RMHMC sampler quickly converges to the true mode after minimal tuning of the integration step-size based on the integration error, which corresponds directly to the acceptance rate. mMALA also converges very quickly to the true posterior mode. MALA converges in a similar number of iterations, but only for a suitable choice of scaling factor. The right-middle plot in Figure 4.5 shows convergence when the scaling factor is carefully tuned for the transient phase of the Markov chain, however the right hand plot demonstrates how it fails to converge at all given a scaling factor which is tuned for stationarity. Detailed results of the sampling efficiency of each method are given in Table 4.10. In this example the RMHMC method required just 1.5 seconds per effectively independent sample compared to more than 2 hours needed by MALA. In addition to taking far longer to sample, MALA also generates much more highly correlated samples and as a result has a far worse effective sample size. This can also be seen in Figure 4.6 which shows the inferred posterior latent field, the posterior latent process and the variance associated with the Monte Carlo estimate. For RMHMC, the variance in the estimates increases where there the data sample is small, i.e. in the top right hand corner of the

Table 4.10: Comparison of sampling methods for the latent variables of a log-Gaussian Cox Process. We note that MALA requires the use of a Cholesky decomposition at every iteration, which scales as $\mathcal{O}(n^3)$. The manifold methods, mMALA and RMHMC, require the inverse of the metric tensor to be calculated, which also scales as $\mathcal{O}(n^3)$, however this need only be calculated once as the metric is constant for the latent variables of this model.

Method	Time	ESS (Min, Med, Max)	s/Min ESS	Rel. Speed
MALA with Trans. (Transient)	31,577	(3, 8, 50)	10,605	$\times 1$
MALA with Trans. (Stationary)	31,118	(4, 16, 80)	7836	$\times 1.35$
mMALA	634	(26, 84, 174)	24.1	$\times 440$
RMHMC	2936	(1951, 4545, 5000)	1.5	$\times 7070$

field. mMALA has slightly more variability, while the low ESS of the MALA methods manifests itself in patchy regions of high variability across the entire field. We note that MALA tuned for stationarity has slightly lower variance than MALA tuned for the transient phase, as one would expect.

Conditionally sampling the hyperparameters using RMHMC proves more costly, with 5000 posterior samples taking around 90 hours of computation time. However, the posterior estimates for the hyperparameters correspond extremely well to their true values, see Figure 4.7.

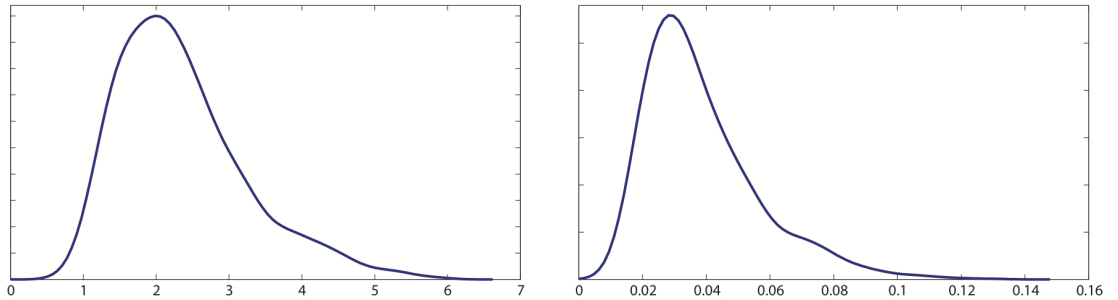


Figure 4.7: Kernel density estimates of the hyperparameters of the log-Gaussian Cox model - The hyperparameter samples were obtained from Gibbs style sampling from the log-Gaussian Cox model. The true values are $\sigma = 1.9$ (left hand plot) and $\beta = 0.03$ (right hand plot).

Inferring the latent field of a log-Gaussian Cox process with a finely grained discretisation is clearly a very challenging problem due to the high dimensionality and

strong spatial correlations present between the latent variables. The major challenges associated with employing MALA are firstly finding a suitable reparameterisation of the target density, and secondly making a suitable choice for the scaling factor according to whether the Markov chain is in a transient or stationary regime. In contrast, mMALA and RMHMC do not exhibit such extreme technical difficulties. We have demonstrated that RMHMC is able to sample the latent variables directly with minimal tuning and effort and without the need for reparameterisation. By employing a Gibbs style sampling scheme we were additionally able to sample the hyperparameters of the covariance function in a relatively computationally efficient manner. An investigation into the sparse approaches presented in (174, 201) may provide further computational efficiencies.

4.5 Conclusions

We have considered a variety of challenging statistical models, whose dimensionality varies from a few parameters to several thousand latent variables. For each of these models we investigated the use of a number of commonly used MCMC methods and compared these with the differential geometric methods we introduced in Chapter 3. In terms of raw ESS, RMHMC consistently offered the best results, often drawing samples with near independence. This generalisation of HMC to a Riemannian manifold can however be computationally costly when the metric is not constant over the parameter space, since it is necessary to calculate 2nd order derivatives of the log-likelihood with respect to the parameters to calculate the derivative of the metric tensor. In addition the generalised leapfrog integrator requires multiple fixed point iterations that can also be costly to compute. Similarly, the full mMALA scheme can also be computationally expensive due to the need to calculate derivatives of the metric tensor, although it does avoid the use of fixed point iterations. In practice the pragmatic approach of employing a constant metric tensor, in RMHMC, and a locally constant metric tensor, in a simplified mMALA scheme, often works most effectively when taking into account computational cost. Such approaches still make use of the local Riemannian geometry and appear to work very well for many challenging models.

In the next chapter we return to the motivating example of statistical models based on systems of differential equations, which are very useful for describing and predicting

complex natural processes occurring in the physical and, in particular, the biological sciences.

5

Statistical Inference over Dynamical Systems

Many dynamical systems in engineering, economics and, in particular, the natural sciences can be described using ordinary differential equations (ODEs). The forward problem consists of numerically obtaining a solution to a system of ODEs given the model parameters and initial conditions, and this may be easily achieved using commonly available software. A much trickier task is the so called inverse problem; can we work out the original parameter values of an ODE given some noisy measurements of its solution? Since the 1970s, many approaches have been proposed to tackle this problem from a mathematical and engineering perspective, during which time computer software became publicly available for the automatic integration of many classes of ODEs (70). This has often been considered from an inverse operator point of view, where the main problem is the ill-conditioning of the problem (193). It is perhaps surprising that there has been relatively little research on *statistical* methods for parameter estimation of ODEs until only very recently. Bayesian approaches in particular provide a very natural way of analysing the uncertainty in descriptions of dynamical systems defined by nonlinear differential equations.

We recall that a dynamical system may be described by a collection of N ordinary differential equations and model parameters $\boldsymbol{\theta}$, which define a functional relationship between the process state, $\mathbf{x}(t)$, and its time derivative such that

$$\dot{\mathbf{x}}(t) \equiv \frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}) \quad (5.1)$$

The first optimisation based approaches to this inverse problem were proposed around 30 years ago and were based on solving the initial value problem (11, 19, 54). Given initial conditions, the system of ODEs can be numerically solved and then compared to the observed data. Using for example a least squares approach, the model parameters can then be optimised. The main concern is usually the computational expense of numerically solving ODE systems for many sets of parameters. The optimisation task is made harder by the fact that nonlinear ODEs generally induce complex objective functions with multiple local minima, given by

$$\sum_{n=1}^N \sum_{t=1}^T (y_n(t) - x_n(t))^2 \quad (5.2)$$

where $x_n(t)$ is the solution for the n th state at time point t to the given system of differential equations, and $y_n(t)$ is the corresponding observed data.

An alternative approach (202) was soon suggested involving a very different two-step procedure in an attempt to circumvent the computational issues. First, a spline approximation was fitted to the data using a least squares approach. Then the parameters of the model were fitted, again using least squares, according to the following alternative objective function,

$$\sum_{n=1}^N \sum_{t=1}^T (f_n(m(t), \theta) - m'_n(t))^2 \quad (5.3)$$

where now $m'_n(t)$ is the derivative estimate from the spline, and $f_n(m(t), \theta)$ is the differential equation function evaluated using the smoothed observation estimates from the spline approximation and the parameter values being optimised. Any optimisation is therefore done over the *derivative* space instead of the state space. This is computationally much faster than explicitly solving the ODE for each set of parameter values and the smoothing of the data also results in a smoother objective function that is easier to optimise, however the drawback is that we must now work with an extended statistical model in which inferences are made only approximately; the parameter values are chosen based on the spline approximation of the data, instead of directly on the data itself.

More recently there has been somewhat of a resurgence of developing these approaches, in particular viewing them from a statistical perspective. Smoothing based methods may be useful for obtaining approximate parameter estimates, however ideally we want to be able to fully quantify the uncertainty associated with both the model parameters and the model itself, so that we may rank multiple hypotheses that encode the possible structures of a system as ODE models. We therefore focus on Bayesian methods for these differential equation models.

In this chapter we consider statistical approaches that allow us to analyse one of the motivating examples mentioned in Chapter 1; using the differential equation formalism we wish to model the biochemical structure responsible for circadian rhythms in plants. There are two main strategies for reducing the cost of performing Bayesian inference in this setting. We can either employ samplers that explore the parameter space more efficiently, thereby increasing the raw ESS, or we can derive alternative sampling schemes based on computationally less expensive approximations such as using an alternative likelihood function; in this chapter we consider both approaches.

We first consider the application of the differential geometric sampling methods developed in Chapter 3 to statistical models based on nonlinear systems of differential equations. In particular we show that we can obtain more efficient sampling of the posterior distribution by employing local sensitivity information from an ODE model, which coincides with employing 2nd order geometric information in the form of the Expected Fisher Information. This section follows from the published paper (77).

Having established efficient manifold MCMC sampling methodology based on explicitly solving systems of ODEs, we take a slight digression to look at an alternative approach to performing inference over such models. We consider an approximate surrogate likelihood approach that avoids explicit solution of the ODE system to see whether this could further improve computational efficiency, which might be useful for the task of analysing the larger, more realistic biological dynamical systems that motivate this work. In particular we investigate the use of Gaussian processes to smooth the data, whilst performing inference on the parameters using the derivative estimates. This approach allows for uncertainty to be propagated from the smoothing step through to the parameter inference step, and we consider both ordinary and delay differential equation models. This section also follows from recently published work (29).

Finally, we examine the modelling issues that arise when using the differential equation formalism by considering some disease outbreak models. The models we use are very simple, however they demonstrate clearly the Bayesian approach in such a setting and we investigate the effects that changes in the priors and data have on the final inferences. This analysis will provide useful insights before tackling the larger, more complex biological systems in Chapter 6.

5.1 Manifold Sampling for ODE Models

We have already seen an example in Chapter 1 of a posterior distribution with strong correlation structure that was induced by a very simple pair of coupled differential equations. For larger, more complex models there may well be highly nonlinear relationships between the inferred parameters, which may make the higher dimensional posterior distribution much harder to sample efficiently from, particularly when using a standard Metropolis-Hastings sampler. In this section we consider the application of the manifold sampling methods, introduced in Chapter 3, to such ODE models. We see that by employing a sampling scheme that explicitly takes into account the first and second order sensitivities of the differential equations in the model, we may obtain samples from the posterior distribution far more efficiently, as the sampler *automatically* takes into account the local geometry at each point in the parameter space.

A dynamical system may be described by a collection of N nonlinear ordinary differential equations and model parameters $\boldsymbol{\theta}$ which define a functional relationship between the process state, $\mathbf{x}(t)$, and its time derivative such that $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta})$. A sequence of process observations, $\mathbf{y}(t)$, are usually assumed to be contaminated with some measurement error, which is modelled as $\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t)$, where $\boldsymbol{\epsilon}(t)$ defines an appropriate multivariate noise process, e.g. a zero-mean Gaussian with variance σ_n^2 for each of the N states. If observations are made at T distinct time points, the $T \times N$ matrices summarise the overall observed system as $\mathbf{Y} = \mathbf{X} + \mathbf{E}$. In order to obtain values for \mathbf{X} , the system of ODEs must be solved, so that in the case of an initial value problem $\mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)$ denotes the solution of the system of equations at the specified time points for the parameters $\boldsymbol{\theta}$ and initial conditions \mathbf{x}_0 . The posterior density follows by employing appropriate priors such that

$$p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{x}_0, \boldsymbol{\sigma}) \propto \pi(\boldsymbol{\theta}) \prod_n \mathcal{N}(\mathbf{Y}_{:,n}|\mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{:,n}, \mathbf{I}_T \sigma_n) \quad (5.4)$$

5.1.1 First Order Sensitivities

If we wish to employ sampling methods involving first order geometric information, such as MALA or HMC, we must be able to calculate the derivatives of the likelihood with respect to each of the model parameters. If the log-likelihood is given by

$$\mathcal{L}(\mathbf{Y}_{:,n}|\mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{:,n}, \boldsymbol{\Sigma}_n) \propto \sum_{n=1}^N -\frac{1}{2} (\mathbf{Y}_{:,n} - \mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{:,n})^T \boldsymbol{\Sigma}_n^{-1} (\mathbf{Y}_{:,n} - \mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{:,n}) \quad (5.5)$$

then the derivative with respect to the i th parameter follows as

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \sum_{n=1}^N (\mathbf{Y}_{:,n} - \mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{:,n})^T \boldsymbol{\Sigma}_n^{-1} \mathbf{S}_{:,n}^i \quad (5.6)$$

where the T -dimensional vectors of first order sensitivities for the n th component of state relative to the i th parameter are denoted as $\mathbf{S}_{:,n}^i = \partial \mathbf{X}_{:,n} / \partial \theta_i$. We see that the first order sensitivities of the differential equations in the model therefore enter into this expression. One method of obtaining these required sensitivities at all time points is to approximate them using finite differences, however this approach is generally not very accurate. For this example we differentiate the system of equations with respect to each of the parameters and directly solve the first order sensitivity equations defined as follows

$$\dot{\mathbf{S}}_{t,n}^i = \frac{\partial \mathbf{f}_n(\mathbf{x}, \boldsymbol{\theta}, t)}{\partial \theta_i} = \sum_{l=1}^N \frac{\partial \mathbf{f}_{t,n}}{\partial x_l} \mathbf{S}_{t,l}^i + \frac{\partial \mathbf{f}_{t,n}}{\partial \theta_i}$$

We note that we must take the *total derivative* with respect to $\boldsymbol{\theta}$, since the states \mathbf{x} also depend on the parameter values. We may augment the original system with these new differential equations, and then numerically integrate this new system with respect to time in order to obtain both the states and the sensitivities of the states. The computational time required to solve these equations will of course be more than

solving the original system, however we note that the additional equations are linear ODEs of the sensitivities, so we might not expect such a large increased computational expenditure.

5.1.2 Second Order Sensitivities

We also need to calculate a metric tensor if we wish to employ manifold sampling methods for this statistical model. By considering the Gaussian noise model described above, where $\Sigma_n = \mathbf{I}_T \sigma_n^2$, using the Expected Fisher Information we straightforwardly obtain the following analytic expressions for the metric tensor and its derivatives in terms of the first and second order sensitivities of the states of the differential equations. This metric tensor and its derivatives follow as

$$\mathbf{G}(\boldsymbol{\theta})_{ij} = \sum_{n=1}^N \mathbf{S}_{:,n}^{i\top} \Sigma_n^{-1} \mathbf{S}_{:,n}^j \quad \frac{\partial \mathbf{G}(\boldsymbol{\theta})_{ij}}{\partial \theta_k} = \sum_{n=1}^N \left(\frac{\partial \mathbf{S}_{:,n}^{i\top}}{\partial \theta_k} \Sigma_n^{-1} \mathbf{S}_{:,n}^j + \mathbf{S}_{:,n}^{i\top} \Sigma_n^{-1} \frac{\partial \mathbf{S}_{:,n}^j}{\partial \theta_k} \right)$$

This expression for the metric tensor has an appealing interpretation in that the actual sensitivity equations of the underlying dynamic system model explicitly enter into the proposal process for our manifold based MCMC scheme. In a similar fashion as before, we may augment the original system with additional equations to solve for the second order sensitivities, which are required for calculating the partial derivatives of the metric tensor with respect to the model parameters. These equations follow as

$$\frac{\partial \dot{\mathbf{S}}_{t,n}^i}{\partial \theta_k} = \sum_{l=1}^N \left[\left(\sum_{m=1}^N \frac{\partial^2 \mathbf{f}_{t,n}}{\partial x_l \partial x_m} \mathbf{S}_{t,m}^k + \frac{\partial^2 \mathbf{f}_{t,n}}{\partial x_l \partial \theta_k} \right) \mathbf{S}_{t,l}^i + \frac{\partial \mathbf{f}_{t,n}}{\partial x_l} \frac{\partial \mathbf{S}_{t,l}^i}{\partial \theta_k} \right] + \sum_{l=1}^N \frac{\partial^2 \mathbf{f}_{t,n}}{\partial \theta_i \partial x_l} \mathbf{S}_{t,l}^k + \frac{\partial^2 \mathbf{f}_{t,n}}{\partial \theta_i \partial \theta_k}$$

We now have everything required to implement RMHMC and mMALA sampling schemes for statistical models defined by systems of nonlinear differential equations.

Interestingly the structure of the equations required for the metric tensor and its derivatives are such that RMHMC can be used to form a parallel tempering or population Monte Carlo scheme where the numerical solution of the sensitivity equations and their derivatives can be used at all tempered posterior distributions defined as

$$p(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{x}_0, \boldsymbol{\sigma}, \beta) \propto \pi(\boldsymbol{\theta}) \prod_n \mathcal{N}(\mathbf{Y}_{:,n} | \mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{:,n}, \Sigma_n)^\beta \quad (5.7)$$

where $0 \leq \beta \leq 1$ (as described at the end of Chapter 3) and the metric is a simple scaling

$$\mathbf{G}(\boldsymbol{\theta}, \beta)_{ij} = \beta \sum_{n=1}^N \mathbf{s}_{:,n}^{i\top} \boldsymbol{\Sigma}_n^{-1} \mathbf{s}_{:,n}^j \quad (5.8)$$

5.1.3 Fitzhugh Nagumo Model

We present results comparing the sampling efficiency for the parameters of the Fitzhugh Nagumo differential equations (159),

$$\dot{V} = c \left(V - \frac{V^3}{3} + R \right), \quad \dot{R} = - \left(\frac{V - a + bR}{c} \right) \quad (5.9)$$

We obtain samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{x}_0, \boldsymbol{\sigma})$, and so in this example $\mathbf{X}_{1,\cdot} = \mathbf{V}$ and $\mathbf{X}_{2,\cdot} = \mathbf{R}$. The sampling schemes we employ are Metropolis-Hastings, MALA and HMC, as described in Chapter 4, as well as the manifold methods mMALA, simplified mMALA and RMHMC. We again compare the simulations by calculating the effective sample size (ESS) normalised by the computational time required to produce the samples.

Before proceeding we require the first and second partial derivatives of the Fitzhugh Nagumo equations in order to calculate the metric tensor and its derivatives. In practice, all these expressions may be obtained automatically using symbolic differentiation.

$$\frac{\partial \dot{V}}{\partial a} = \frac{\partial \dot{V}}{\partial b} = 0, \quad \frac{\partial \dot{V}}{\partial c} = \left(V - \frac{V^3}{3} + R \right), \quad \frac{\partial \dot{R}}{\partial a} = \frac{1}{c}, \quad \frac{\partial \dot{R}}{\partial b} = \frac{-R}{c}, \quad \frac{\partial \dot{R}}{\partial c} = \left(\frac{V - a + bR}{c^2} \right)$$

All of the second derivatives of \dot{V} with respect to the model parameters are equal to zero, and the five non-zero second partial derivatives of \dot{R} are as follows,

$$\frac{\partial^2 \dot{R}}{\partial a \partial c} = -\frac{1}{c^2}, \quad \frac{\partial^2 \dot{R}}{\partial b \partial c} = \frac{R}{c^2}, \quad \frac{\partial^2 \dot{R}}{\partial c \partial a} = -\frac{1}{c^2}, \quad \frac{\partial^2 \dot{R}}{\partial c \partial b} = \frac{R}{c^2}, \quad \frac{\partial^2 \dot{R}}{\partial c^2} = 2 \left(\frac{-V + a - bR}{c^3} \right)$$

In addition, the second partial derivatives with respect to all states and parameters are required for writing the differential equation describing the second order sensitivities.

There are again five non-zero second partial derivatives with respect to the states and parameters as follows

$$\frac{\partial^2 \dot{V}}{\partial V \partial c} = 1 - V^2, \quad \frac{\partial^2 \dot{V}}{\partial R \partial c} = 1, \quad \frac{\partial^2 \dot{R}}{\partial V \partial c} = \frac{1}{c^2}, \quad \frac{\partial^2 \dot{R}}{\partial R \partial b} = -\frac{1}{c}, \quad \frac{\partial^2 \dot{R}}{\partial R \partial c} = \frac{b}{c^2}$$

5.1.4 Experimental Results

We used 200 data points generated from the Fitzhugh Nagumo ODE model between $t = 0$ and $t = 20$ with the model parameters $a = 0.2$, $b = 0.2$, $c = 3$ and initial conditions $V(0) = -1$ and $R(0) = 1$. Gaussian distributed noise with standard deviation equal to 0.5 was then added to the data, see Figure 5.1.

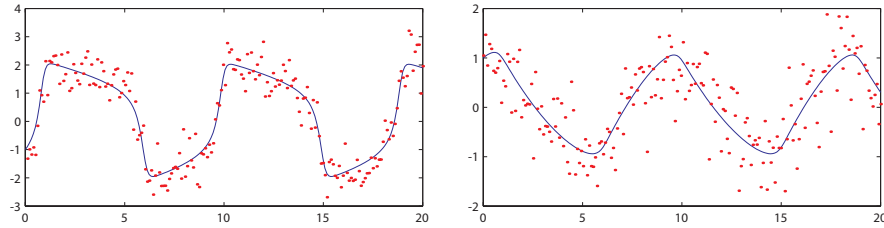


Figure 5.1: Fitzhugh Nagumo ODE model output. - Output for species V (left) and species R (right) of the Fitzhugh Nagumo model with parameters $a = 0.2$, $b = 0.2$, $c = 3$. An example noisy dataset is shown by the red points.

Nonlinear ODEs generally induce corresponding nonlinearities in the target distribution, which may result in many local maxima (27). Careful attention must therefore be paid so that the Markov chains do not converge to the wrong mode, but rather sample from the correct distribution. All the sampling methods employed in this section may be embedded within a population MCMC framework to allow full exploration of and convergence to the target density (27), however for the purpose of comparing sampling efficiency we employ a single Markov chain initialised on the true mode. We collected 5000 posterior samples and calculated the ESS for each parameter, using the minimum value to calculate the time per effectively independent sample. 10 simulations were run for each method, using the same dataset, and all methods were implemented in the interpreted language Matlab for consistency of comparison.

The results of our simulations are shown in Table 5.1. Standard HMC takes the longest time for this problem due to the large number of leapfrog steps it needs to

traverse the parameter space. RMHMC on the other hand requires relatively few leapfrog steps, as it takes into account the local geometry to make better moves. We note however the additional computational cost of the leapfrog steps, during each of which it is necessary to solve the system of ODEs to evaluate the gradients and metric tensor. The first momentum update of RMHMC is relatively quick since only a vector-matrix multiplication is necessary, however updating the parameter values requires the metric tensor to be evaluated for each fixed point iteration in the Generalised leapfrog algorithm as the parameter values converge, thus adding a considerable amount of computation to the overall algorithm. The mMALA methods offer the best performance for this particular example, as they have the benefit of using manifold information to guide the direction of the chain, but without the required fixed point iterations thus only requiring the ODEs to be numerically solved once per iteration. This suggests that mMALA is perhaps particularly suited for settings in which there is a non-constant metric tensor which is expensive to compute, as in this case.

The Fitzhugh Nagumo model has only three parameters and we see that MALA and HMC perform adequately in this low dimensional setting, indeed the largest marginal parameter variance is only four times larger than the smallest marginal variance. We would expect MALA and HMC to perform worse in cases where there is a greater difference in the marginal variances, since the step size of each is restricted by the smallest marginal variance. Similarly, while component-wise Metropolis performs adequately in this setting, we would expect its performance to deteriorate in higher dimensions, as has been previously noted in the literature, when there are greater correlations in the parameters. We shall see an example of this in Chapter 6.

The underlying forward problem of solving initial-value ODEs numerically is a well-studied topic which has spawned very sophisticated numerical solvers. However, in an MCMC context the computational expense required to simulate the ODEs is particularly important, since a very large number of solves may be required to obtain sufficient posterior samples. In the next section we therefore digress to investigate an approximate inference method designed to further reduce the computational cost of inference under certain conditions.

5.2 Gaussian Processes for Approximate ODE Inference

Table 5.1: Summary of results for the Fitzhugh Nagumo model with 10 runs of the parameter sampling scheme and 5000 posterior samples

Sampling Method	Time (s)	Mean ESS (a, b, c)	Total Time/ (Min mean ESS)	Relative Speed
Metropolis	18.5	132, 130, 108	0.17	$\times 3.9$
MALA	14.4	125, 21, 46	0.67	$\times 1$
HMC	815	4668, 3483, 3811	0.23	$\times 2.9$
mMALA	34.9	1057, 925, 956	0.037	$\times 18.1$
mMALA Simp.	14.9	1007, 479, 762	0.031	$\times 21.6$
RMHMC	266	4302, 4202, 3199	0.083	$\times 8$

5.2 Gaussian Processes for Approximate ODE Inference

In the MCMC approaches of the previous section most of the computation time is spent obtaining solutions to the ODEs by solving the forward problem given some parameters and initial conditions. One approach to minimising this cost is by sampling from the space as efficiently as possible, using the local geometric information at each point to make larger proposed steps that are accepted with high probability. An alternative approach is to improve efficiency by replacing the likelihood function with a surrogate that is computationally less expensive to evaluate. We therefore now divagate slightly from manifold sampling to investigate alternatives to explicitly solving the systems of ODEs, which might potentially increase even further the efficiency and speed of statistically analysing the large dynamical systems we shall consider in Chapter 6. In particular we consider an approximate Bayesian method for inferring the parameters in systems of ODEs based on a smoothed approximation to the data and its derivatives using Gaussian processes.

The basic idea is that we first smooth the data to obtain estimates of the underlying state and its derivatives. The model parameters are then fitted to the ODE model based on these state approximations. This two step procedure has previously appeared in the literature using a smoothing spline approach, which we now briefly review. It is of course *assumed* that the spline can accurately describe the dynamics of the data such that errors in the parameter estimates will be minimal, and this is often dependent on having enough accurate data. In particular, parameter values should be checked

5.2 Gaussian Processes for Approximate ODE Inference

a posteriori by solving the ODE model explicitly. An cautionary example is given in (202), showing how parameter values based on a plausible looking, but overfitted, spline approximation results in parameter values whose exact explicit solution is actually quite different from the data. An additional challenge with this approximation approach regards dealing with unobserved species. Varah (202) suggests an example in which the system can be converted into a higher order system of only the observed species, however it is then necessary to make use of 2nd order derivative estimates from the spline, which will likely be subject to even more error than the 1st order estimates.

The two-step method has been built upon recently by the Iterated Principal Differential Analysis framework (157, 160), which iterates between fitting the spline to the data and ensuring fidelity to the ODE model through the penalty term,

$$\sum_{m=1}^M \left(\frac{\partial \hat{\mathbf{x}}_m}{\partial t} - \mathbf{f}(\hat{\mathbf{x}}_m, \boldsymbol{\theta}) \right)^2 \quad (5.10)$$

where this term is based on a number of chosen collocation points at time points indexed by m , which may be larger than the number of data points. We note that as the number of collocation points tends to infinity and this residual tends to zero, we recover the continuous integrated solution of the ODE model. Indeed for large models, the number of collocation points to be optimised may have a large impact on the computational speed of the method. In effect, both states and model parameters must be optimised. More recently, Ramsay et al. (159) introduced the Generalised Smoothing approach with a smoothing parameter λ that offers a way of interpolating between data fit and model fit, with the method being equivalent to a standard least squares model fit to the data as $\lambda \rightarrow \infty$. While this is certainly an advance on the original methods of (202), there are still a number of significant shortcomings. In particular, questions arise concerning the choice of spline parameters that control the smoothing, the choice of interpolation parameter λ , and how these choices ultimately affect the end results.

The methods are all critically dependent on additional regularisation parameters to determine the level of data smoothing. They all exhibit the potential problem of providing sub-optimal point estimates; even (159) may not converge to a reasonable solution depending on the initial values selected, due to the nonlinearity of the optimisation space. Finally, these methods only provide point estimates of the “best”

parameter set and are unable to cope with multiple possible solutions, although it should be noted that (159) does offer a local estimate of uncertainty based on second derivatives computed at a point and at additional computational cost.

5.2.1 Overview

Bayesian, and indeed non-Bayesian, approaches for parameter estimation and model comparison (203) involve evaluating likelihood functions that generally require the explicit numerical solution of a system of differential equations. Similar to ODEs, delay differential equations (DDEs) can also be used to describe certain dynamic systems, where now an explicit time-delay τ is employed. The computational cost of obtaining the required numerical solutions of the ODEs, and in particular DDEs, can result in slow running times.

We now present a two-step method for performing Bayesian inference over mechanistic models by using Gaussian processes (GP) to predict the state variables of the model as well as their derivatives, which avoids the need to solve the system explicitly. In certain cases, this can result in dramatically improved computational efficiency, however this speed increase is not guaranteed in general due to the poor computational scaling of GPs with respect to the number of observations. An additional shortcoming that we must be wary of is that this approach introduces another layer of complexity into our statistical model and inferences will be based on smoothed approximations rather than the data directly, as we discuss later in this section.

We note that state space models offer an alternative approach for performing parameter inference over dynamical models particularly for on-line analysis of data, see (55). Related to the work we present, we also note that in (78) the use of GPs has been proposed in obtaining the solution of fully parameterised linear operator equations such as ODEs. A little known paper by Skilling (181) is also of relevance, in which he suggests that differential equations should be treated as just another inference problem and he describes a method making use of GPs to approximate their solution, however it ultimately fails to be computationally feasible in practice due to the poor computational scaling properties. In a slightly different context, GPs are employed in (163) as emulators of the posterior response to parameter values as a means of improving the computational efficiency of a hybrid Monte Carlo sampler.

We now suggest a Bayesian method that utilises GPs to approximate the solution states of our differential equation model. We demonstrate its speed and statistical accuracy for performing statistical inference over ordinary and delay differential equations under certain conditions, and provide comparisons with the alternative approaches. We also present an example with unobserved species.

5.2.2 Introduction to Gaussian Processes

A Gaussian process (GP) can be considered a natural extension of a Gaussian distribution that is fully defined by a mean function and a covariance function. A GP may be defined over arbitrary dimensions, however since we are only interested in time series we introduce GPs in terms of a one dimensional input space \mathbf{t} , representing time. Both the mean and covariance functions in turn depend on some real process $x(\mathbf{t})$ such that

$$\mu(\mathbf{t}) = E[x(\mathbf{t})] \quad (5.11)$$

$$k(t_1, t_2) = E[(x(t_1) - \mu(t_1))(x(t_2) - \mu(t_2))] \quad (5.12)$$

A GP is therefore a collection of random variables, whose most important property is that any finite number of these random variables have a Gaussian joint distribution. This marginalisation property allows us to perform straightforward computations with this otherwise infinite dimensional object.

As a first example, let us assume we have noise free observations \mathbf{x} at time points \mathbf{t} , and we wish to predict the values \mathbf{x}^* at time points \mathbf{t}^* . Given the assumption that the joint distribution of \mathbf{x} and \mathbf{x}^* is Gaussian distributed, the values of \mathbf{x}^* at \mathbf{t}^* can be calculated using a standard formula. If the joint distribution is given in the standard form

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{x}^* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{x}^*} \end{bmatrix}, \begin{bmatrix} k(\mathbf{t}, \mathbf{t}) & k(\mathbf{t}, \mathbf{t}^*) \\ k(\mathbf{t}^*, \mathbf{t}) & k(\mathbf{t}^*, \mathbf{t}^*) \end{bmatrix} \right) \quad (5.13)$$

then the conditional distribution for \mathbf{x}^* is given by

$$p(\mathbf{x}^* | \mathbf{x}, \mathbf{t}, \mathbf{t}^*) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}^*} + k(\mathbf{t}^*, \mathbf{t})k(\mathbf{t}, \mathbf{t})^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}), k(\mathbf{t}^*, \mathbf{t}^*) - k(\mathbf{t}^*, \mathbf{t})k(\mathbf{t}, \mathbf{t})^{-1}k(\mathbf{t}^*, \mathbf{t})^T)$$

5.2 Gaussian Processes for Approximate ODE Inference

We note that our covariance function defines the type of functions that our GP is capable of describing, for example quickly or slowly varying functions, oscillatory functions etc. Many examples of valid covariance functions are given in (164). Such covariance functions often have hyperparameters defining these characteristics, and we can either fix these if we know exactly what type of functions we want, or we can infer these hyperparameters from the data.

In practice we often assume a zero mean and as we will be dealing with noisy observations we can include an additional term to assume additive independent identically distributed noise, such that our observations take the form $\mathbf{y} = \mathbf{x} + \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$. We can conveniently include this in our covariance function such that our joint distribution then becomes

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{x}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{t}, \mathbf{t}) + \mathbf{I}\sigma^2 & k(\mathbf{t}, \mathbf{t}^*) \\ k(\mathbf{t}^*, \mathbf{t}) & k(\mathbf{t}^*, \mathbf{t}^*) \end{bmatrix}\right) \quad (5.14)$$

and given hyperparameters ϕ and σ we can draw smooth functions from the marginal distribution

$$p(\mathbf{x}^* | \mathbf{y}, \mathbf{t}, \mathbf{t}^*, \sigma, \phi) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (5.15)$$

where ϕ are the hyperparameters of the GP and

$$\boldsymbol{\mu} = k(\mathbf{t}, \mathbf{t}^*)[k(\mathbf{t}, \mathbf{t}) + \mathbf{I}\sigma^2]^{-1}\mathbf{y} \quad (5.16)$$

$$\boldsymbol{\Sigma} = k(\mathbf{t}^*, \mathbf{t}^*) - k(\mathbf{t}, \mathbf{t}^*)[k(\mathbf{t}, \mathbf{t}) + \mathbf{I}\sigma^2]^{-1}k(\mathbf{t}, \mathbf{t}^*)^T \quad (5.17)$$

Furthermore, by noticing that $\mathbf{y} \sim \mathcal{N}(k(\mathbf{t}, \mathbf{t}) + \mathbf{I}\sigma^2)$, we can also sample the hyperparameters of our GP by putting priors over σ and ϕ and sampling from

$$p(\phi, \sigma) \propto p(\mathbf{y} | \phi, \sigma)p(\phi)p(\sigma) \quad (5.18)$$

and so we obtain an expression for the full joint posterior $p(\mathbf{x}^*, \sigma, \phi | \mathbf{y}, \mathbf{t}, \mathbf{t}^*)$. We note that a non-Gaussian noise model may alternatively be implemented using warped GPs (183).

5.2 Gaussian Processes for Approximate ODE Inference

Another useful feature is that we can analytically obtain derivatives of functions sampled from our GP for particular classes of covariance functions - all we require is that the covariance function can be differentiated with respect to each of the inputs, \mathbf{t}_1 and \mathbf{t}_2 in our case. Letting k' , $'k$ and k'' be our covariance function differentiated with respect to the first, second and both input variables \mathbf{t}_1 and \mathbf{t}_2 respectively, the observed data and derivative predictions are again jointly Gaussian distributed (184),

$$\begin{bmatrix} \mathbf{y} \\ \dot{\mathbf{x}}^* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{t}, \mathbf{t}) + \mathbf{I}\sigma^2 & k'(\mathbf{t}, \mathbf{t}^*) \\ 'k(\mathbf{t}^*, \mathbf{t}) & k''(\mathbf{t}^*, \mathbf{t}^*) \end{bmatrix} \right) \quad (5.19)$$

and so derivative predictions follow in a similar manner as before,

$$p(\dot{\mathbf{x}}^* | \mathbf{y}, \mathbf{t}, \mathbf{t}^*, \sigma, \phi) \sim \mathcal{N}(\mathbf{m}, \mathbf{K}) \quad (5.20)$$

where

$$\mathbf{m} = 'k(\mathbf{t}, \mathbf{t}^*)[k(\mathbf{t}, \mathbf{t}) + \mathbf{I}\sigma^2]^{-1}\mathbf{y} \quad (5.21)$$

$$\mathbf{K} = k''(\mathbf{t}^*, \mathbf{t}^*) - 'k(\mathbf{t}, \mathbf{t}^*)[k(\mathbf{t}, \mathbf{t}) + \mathbf{I}\sigma^2]^{-1}'k'(\mathbf{t}, \mathbf{t}^*)^T \quad (5.22)$$

We therefore have a Bayesian nonparametric smoothing of our data, for which we can easily calculate derivatives, with uncertainty estimates.

5.2.3 Auxiliary Gaussian Processes on State Variables

Returning to the problem of inferring parameters in an ODE model, we see that we could produce a two step approach, whereby we employ a likelihood function based on the mismatch between the smoothed derivative estimates from the GP and the ODE function output based on the parameters and the smoothed data estimates from the GP.

Since the differential equations we consider are reasonably smooth and well-behaved, we employ a standard squared exponential covariance function (164), which is infinitely differentiable. We employ a simple zero mean Gaussian process prior since the shape of the GP is inferred directly from the data, with the model solutions subsequently being enforced through the likelihood function. We note that an alternative approach is to

5.2 Gaussian Processes for Approximate ODE Inference

encode the ODE model directly into the covariance function of the Gaussian process, as has been done for Latent Force models (4), instead of separating the data driven and mechanistic modelling components, as is the case in this work (see Figure 5.2(b)).

In particular we obtain a posterior over the model parameters by assuming Normal errors between the derivatives $\dot{\mathbf{x}}_{n,\cdot}$, for each of the N states of our ODE model, and the functional $\mathbf{f}_n(\mathbf{x}, \boldsymbol{\theta}, \mathbf{t})$, which is estimated using the state predictions from the GP at each of the specified time points. Then $p(\dot{\mathbf{x}}_{n,\cdot} | \mathbf{x}, \boldsymbol{\theta}, \gamma_n) = \mathcal{N}_{\dot{\mathbf{x}}_{n,\cdot}}(\mathbf{f}_n(\mathbf{x}, \boldsymbol{\theta}, \mathbf{t}), \mathbf{I}\gamma_n)$, where we now have error variances γ_n and σ_n for each state.

We therefore have two statistical models describing the state derivatives; one in terms of the data and one in terms of the ODE model. We can choose to model these jointly as a simple product such that

$$p(\dot{\mathbf{x}}_{n,\cdot} | \mathbf{x}, \boldsymbol{\theta}, \gamma_n, \phi_n, \sigma_n) \propto \mathcal{N}_{\dot{\mathbf{x}}_{n,\cdot}}(\mathbf{m}_n, \mathbf{K}_n) \mathcal{N}_{\dot{\mathbf{x}}_{n,\cdot}}(\mathbf{f}_n(\mathbf{x}, \boldsymbol{\theta}, \mathbf{t}), \mathbf{I}\gamma_n) \quad (5.23)$$

Finally, including priors $\pi(\boldsymbol{\theta})$ and $\pi(\gamma) = \prod_n \pi(\gamma_n)$ we obtain

$$\begin{aligned} p(\boldsymbol{\theta}, \gamma | \mathbf{X}, \boldsymbol{\varphi}, \boldsymbol{\sigma}) &= \int p(\dot{\mathbf{X}}, \boldsymbol{\theta}, \gamma | \mathbf{X}, \boldsymbol{\varphi}, \boldsymbol{\sigma}) d\dot{\mathbf{X}} \\ &\propto \pi(\boldsymbol{\theta}) \pi(\gamma) \prod_n \int \mathcal{N}(\mathbf{m}_n, \mathbf{K}_n) \mathcal{N}(\mathbf{f}_n(\mathbf{X}, \boldsymbol{\theta}, \mathbf{t}), \mathbf{I}\gamma_n) d\dot{\mathbf{X}}_{n,\cdot} \\ &\propto \frac{\pi(\boldsymbol{\theta}) \pi(\gamma)}{\prod_n \mathcal{Z}(\gamma_n)} \exp \left\{ -\frac{1}{2} \sum_n (\mathbf{f}_n - \mathbf{m}_n)^\top (\mathbf{K}_n + \mathbf{I}\gamma_n)^{-1} (\mathbf{f}_n - \mathbf{m}_n) \right\} \end{aligned}$$

where $\mathbf{f}_n \equiv \mathbf{f}_n(\mathbf{X}, \boldsymbol{\theta}, \mathbf{t})$, and $\mathcal{Z}(\gamma_n) = |2\pi(\mathbf{K}_n + \mathbf{I}\gamma_n)|^{\frac{1}{2}}$ is a normalising constant. The gradients can be marginalised exactly and we obtain a straightforward method for sampling $\boldsymbol{\theta}$

5.2.4 Sampling Schemes for Fully Observed Systems

The introduction of the auxiliary model and its associated variables has enabled us to recast the differential equation as another component of the inference process. The relationship between the auxiliary model and the physical process that we are modelling is shown in Figure 5.2(b), where the dotted lines represent a transfer of information between the models. This information transfer takes place through sampling candidate

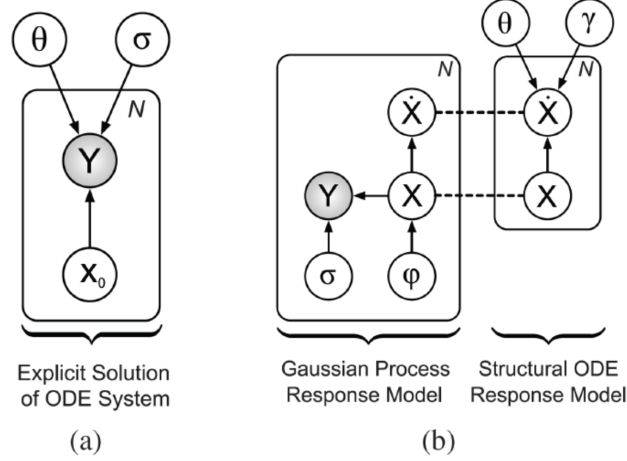


Figure 5.2: Graphical models representing different approaches to inference over differential equation systems. - (a) Graphical model representing explicit solution of an ODE system, (b) Graphical model representing approach developed in this chapter with dashed lines showing how the two models are combined in product form.

solutions for the system in the GP model. Inference is performed by combining these approximate solutions with the system dynamics from the differential equations. It now remains to define an overall sampling scheme for the structural parameters. We assume that the system is defined in terms of ODEs, however we note that our scheme is easily extended for delay differential equations (DDEs), where now predictions at each time point t and the associated delay $(t - \tau)$ are required. Predictions of these additional time points can be easily obtained with the prediction inputs $(t - \tau)$. We present results for a DDE system in Section 5.2.7. We can now consider the complete sampling scheme by also inferring the hyperparameters and corresponding predictions of the state variables and derivatives using the GP framework described in Section 5.2.3. We can obtain samples θ from the desired marginal posterior $p(\theta|\mathbf{Y})$ ¹ by sampling from the joint posterior $p(\theta, \gamma, \mathbf{X}, \varphi, \sigma|\mathbf{Y})$ as follows

¹Note that this is implicitly conditioned on the class of covariance function chosen.

$$p(\boldsymbol{\varphi}_n, \sigma_n | \mathbf{Y}_{n,\cdot}) \propto \pi(\sigma_n) \pi(\boldsymbol{\varphi}_n) \mathcal{N}_{\mathbf{Y}_{n,\cdot}}(\mathbf{0}, \sigma_n^2 \mathbf{I} + \mathbf{C}_{\boldsymbol{\varphi}_n}) \quad (5.24)$$

$$p(\mathbf{X}_{n,\cdot} | \mathbf{Y}_{n,\cdot}, \sigma_n, \boldsymbol{\varphi}_n) = \mathcal{N}_{\mathbf{X}_{n,\cdot}}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \quad (5.25)$$

$$p(\gamma | \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\sigma}) \propto \frac{\pi(\gamma)}{\prod_n \mathcal{Z}(\gamma_n)} \exp \left\{ -\frac{1}{2} \sum_n \boldsymbol{\delta}_n^\top (\mathbf{K}_n + \mathbf{I}_{\gamma_n})^{-1} \boldsymbol{\delta}_n \right\} \quad (5.26)$$

$$p(\boldsymbol{\theta} | \mathbf{X}, \gamma, \boldsymbol{\varphi}, \boldsymbol{\sigma}) \propto \pi(\boldsymbol{\theta}) \exp \left\{ -\frac{1}{2} \sum_n \boldsymbol{\delta}_n^\top (\mathbf{K}_n + \mathbf{I}_{\gamma_n})^{-1} \boldsymbol{\delta}_n \right\} \quad (5.27)$$

where $\boldsymbol{\delta}_n \equiv \mathbf{f}_n - \mathbf{m}_n$. This requires two sampling schemes; one for inferring the parameters of the GP, $\boldsymbol{\varphi}$ and $\boldsymbol{\sigma}$, and another for the parameters of the structural system, $\boldsymbol{\theta}$ and γ . As we are also sampling the hyperparameters, we can marginalise these out, reducing the potential problem of overfitting, as was demonstrated previously by Varah (202).

Figure 5.2(a) illustrates graphically the conditional dependencies of the overall statistical model and from this the posterior density follows by employing appropriate priors such that $p(\boldsymbol{\theta}, \mathbf{x}_0, \boldsymbol{\sigma} | \mathbf{Y}) \propto \pi(\boldsymbol{\theta}) \pi(\mathbf{x}_0) \pi(\boldsymbol{\sigma}) \prod_n \mathcal{N}_{\mathbf{Y}_{n,\cdot}}(\mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{n,\cdot}, \mathbf{I} \sigma_n^2)$. The desired marginal $p(\boldsymbol{\theta} | \mathbf{Y})$ can be obtained from this joint posterior¹.

Using an explicit approach, we can run into difficulties sampling from multimodal posteriors. Recent advances in MCMC methodology suggest solutions to this problem in the form of population-based MCMC methods (96), which we therefore implement to sample the structural parameters of our model. An advantage of the auxiliary variable smoothing approach is that the posterior distribution is also smoothed, allowing for easier sampling as noted by Ramsay et al. (159).

Sampling of the GP covariance function parameters requires computation of a matrix determinant and its inverse, so for all N states in the system a dominant scaling of $\mathcal{O}(NT^3)$ will be obtained. This poses little problem for many applications in systems biology since T is often fairly small ($T \approx 10$ to 100). For larger values of T , sparse approximations can offer much improved computational scaling of order $\mathcal{O}(NM^2T)$, where M is the number of time points selected (120). Sampling from a multivariate Normal whose covariance matrix and corresponding decompositions have already been computed therefore incurs no dominating additional computational overhead. The final

¹This distribution is implicitly conditioned on the numerical solver and associated error tolerances.

sampling step requires each of the \mathbf{K}_n matrices to be constructed, thus incurring a total $\mathcal{O}(NT^3)$ scaling per sample.

An approximate scheme can be constructed by first obtaining the *maximum a posteriori* values for the GP hyperparameters and posterior mean state values, $\hat{\boldsymbol{\varphi}}$, $\hat{\boldsymbol{\sigma}}$, $\hat{\mathbf{X}}_n$, and then employing these in Equation 5.27. This will provide samples from $p(\boldsymbol{\theta}, \gamma | \hat{\mathbf{X}}, \hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\sigma}}, \mathbf{Y})$ which may be a useful surrogate for the full joint posterior incurring lower computational cost as all matrix operations will have been pre-computed.

5.2.5 Extension to Partially Observed Systems

We can also construct a sampling scheme for the special case where some states are unobserved. We partition \mathbf{X} into \mathbf{X}_o , and \mathbf{X}_u . Let o index the observed states, then we may infer all the unknown variables as follows

$$p(\boldsymbol{\theta}, \gamma, \mathbf{X}_u | \mathbf{X}_o, \boldsymbol{\varphi}, \boldsymbol{\sigma}) \propto \pi(\boldsymbol{\theta})\pi(\gamma)\pi(\mathbf{X}_u) \exp \left\{ -\frac{1}{2} \sum_{n \in o} (\boldsymbol{\delta}_n^{o,u})^\top (\mathbf{K}_n + \mathbf{I}\gamma_n)^{-1} (\boldsymbol{\delta}_n^{o,u}) \right\}$$

where $\boldsymbol{\delta}_n^{o,u} \equiv \mathbf{f}_n(\mathbf{X}_o, \mathbf{X}_u, \boldsymbol{\theta}, \mathbf{t}) - \mathbf{m}_n$ and $\pi(\mathbf{X}_u)$ is an appropriately chosen prior. The values of unobserved species are obtained by propagating their sampled initial values using the corresponding discrete versions of the differential equations and the smoothed estimates of observed species. The p53 transcriptional network example we consider shortly requires inference over unobserved protein species, see Section 5.2.8. We now demonstrate our GP-based inference method using a standard squared exponential covariance function on a variety of examples involving both ordinary and delay differential equations, and compare the accuracy and speed.

5.2.6 Example 1 - Nonlinear Ordinary Differential Equations

We first consider the Fitzhugh Nagumo model (159) which was originally developed to model the behaviour of spike potentials in the giant axon of squid neurons and is defined as

$$\frac{dV}{dt} = c \left(V - \frac{V^3}{3} + R \right) \quad \frac{dR}{dt} = \frac{1}{c} (V - a + bR)$$

5.2 Gaussian Processes for Approximate ODE Inference

Although consisting of only 2 equations and 3 parameters, this dynamical system exhibits a highly nonlinear likelihood surface (159), which is induced by the sharp changes in the properties of the limit cycle as the values of the parameters vary. Such a feature is common to many nonlinear systems and so this model provides an excellent test for our GP-based parameter inference method.

Data is generated from the model, with parameters $a = 0.2$, $b = 0.2$, $c = 3$, at $\{40, 80, 120\}$ time points with additive Gaussian noise, $N(0, v)$ for $v = 0.1 \times \sigma_n$, where σ_n is the standard deviation for the n th species. The parameters were then inferred from these data sets using the full Bayesian sampling scheme and the approximate sampling scheme, both employing population MCMC. Additionally, we inferred the parameters using 2 alternative methods, the profiled estimation method of Ramsay et al. (159) and a population MCMC sampling scheme, in which the ODEs were solved explicitly, to complete the comparative study.

Fitzhugh Nagumo ODE Model				
Samples	Method	a	b	c
40	GP MAP	0.1930 ± 0.0242	0.2070 ± 0.0453	2.9737 ± 0.0802
	GP Fully Bayesian	0.1983 ± 0.0231	0.2097 ± 0.0481	3.0133 ± 0.0632
	Explicit ODE	0.2015 ± 0.0107	0.2106 ± 0.0385	3.0153 ± 0.0247
80	GP MAP	0.1950 ± 0.0206	0.2114 ± 0.0386	2.9801 ± 0.0689
	GP Fully Bayesian	0.2068 ± 0.0194	0.1947 ± 0.0413	3.0139 ± 0.0585
	Explicit ODE	0.2029 ± 0.0121	0.1837 ± 0.0304	3.0099 ± 0.0158
120	GP MAP	0.1918 ± 0.0145	0.2088 ± 0.0317	3.0137 ± 0.0489
	GP Fully Bayesian	0.1971 ± 0.0162	0.2081 ± 0.0330	3.0069 ± 0.0593
	Explicit ODE	0.2071 ± 0.0112	0.2123 ± 0.0286	3.0112 ± 0.0139

Table 5.2: Summary statistics for each of the inferred parameters of the Fitzhugh Nagumo model

All the algorithms were coded in Matlab, and the population MCMC algorithms were run with 30 temperatures, and used a suitably diffuse $\Gamma(2, 1)$ prior distribution for all parameters, forming the base distribution for the sampler. Two of these population MCMC samplers were run in parallel and the \hat{R} statistic (72) was used to monitor convergence of all chains at all temperatures. The required numerical approximations to the ODE were calculated using the Sundials ODE solver (90, 178), which has been demonstrated to be considerably (up to 100 times) faster than the standard

5.2 Gaussian Processes for Approximate ODE Inference

ODE45/ODE15s solvers commonly used in Matlab. In our experiments the chains generally converged after around 5000 iterations, and 2000 samples were then drawn to form the posterior distributions. Ramsay’s method (159) was implemented using the Matlab code which accompanies their paper. The optimal algorithm settings were used, tuned for the Fitzhugh Nagumo model (see (159) for details) which they also investigated. Each experiment was repeated 100 times, and Table 1 shows summary statistics for each of the inferred parameters. All of the three sampling methods based on population MCMC produced low variance samples from posteriors positioned close to the true parameters values. Most noticeable from the results in Figure 5.3 is the speed advantage the GP based methods have over the more direct approach, whereby the differential equations are solved explicitly; the GP methods introduced in this chapter offer up to a 10-fold increase in speed, even for this relatively simple system of ODEs.

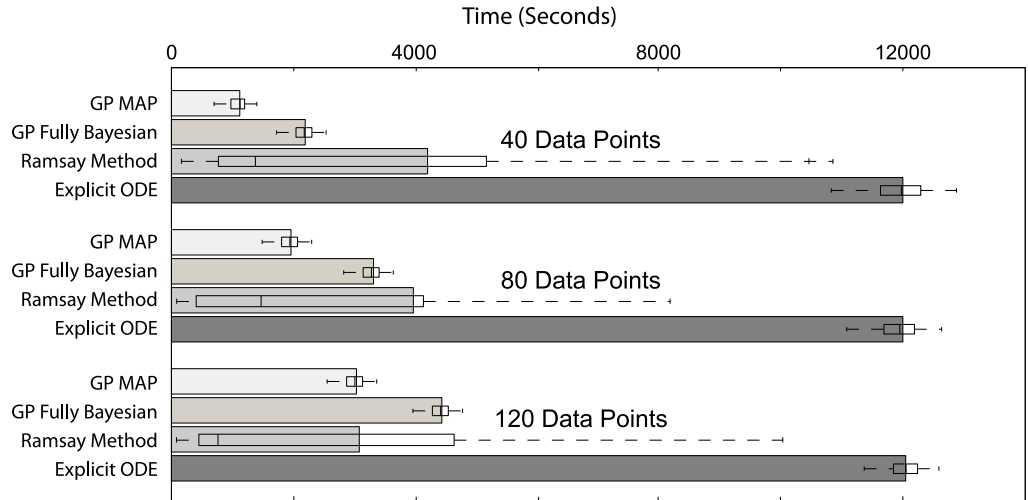


Figure 5.3: Summary statistics for time taken to perform ODE inference using a variety of methods. - Summary statistics of the overall time taken for the algorithms to run to completion. Solid bars show mean time for 100 runs; superimposed boxplots display median results with upper and lower quartiles.

We found the performance of the profiled estimation method (159) to be very sensitive to the initial parameter values. In practice parameter values are unknown, indeed little may be known even about the range of possible values they may take, thus it seems sensible to choose initial values from a wide prior distribution so as to explore as many regions of parameter space as possible. Employing profiled es-

timation using initial parameter values drawn from a wide gamma prior, however, yielded highly biased results with the algorithm often converging to local maxima far from the true parameter values. The parameter estimates become more biased as the variance of the prior is increased and as the starting points move further from the true parameter values. Consider parameter a ; for 40 data points, for initial values $a, b, c \sim \mathcal{N}(\{0.2, 0.2, 3\}, 0.2)$, the range of estimated values for \hat{a} was $[\text{Min}, \text{Median}, \text{Max}] = [0.173, 0.203, 0.235]$. For initial values $a, b, c \sim \Gamma(1, 0.5)$, the \hat{a} had a range $[\text{Min}, \text{Median}, \text{Max}] = [-0.329, 0.205, 9.3 \times 10^9]$ and for a wider prior $a, b, c \sim \Gamma(2, 1)$, then \hat{a} had range $[\text{Min}, \text{Median}, \text{Max}] = [-1.4 \times 10^{10}, 0.195, 2.2 \times 10^9]$. Lack of robustness therefore seems to be a significant problem with this profiled estimation method. The speed of the profiled estimation method was also extremely variable, and this was observed to be very dependent on the initial parameter values. For initial values $a, b, c \sim \mathcal{N}(\{0.2, 0.2, 3\}, 0.2)$, the times recorded were $[\text{Min}, \text{Mean}, \text{Max}] = [193, 308, 475]$. Using initial values sampled from a different prior, such that $a, b, c \sim \Gamma(1, 0.5)$, the times were $[\text{Min}, \text{Mean}, \text{Max}] = [200, 913, 3265]$ and similarly for a wider prior $a, b, c \sim \Gamma(2, 1)$, $[\text{Min}, \text{Mean}, \text{Max}] = [132, 4171, 37411]$.

5.2.7 Example 2 - Nonlinear Delay Differential Equations

This example model describes the oscillatory behaviour of the concentration of mRNA and its corresponding protein level in a genetic regulatory network, introduced by Monk (140). The translocation of mRNA from the nucleus to the cytosol is explicitly described by a delay differential equation,

$$\frac{d\mu}{dt} = \frac{1}{1 + (p(t - \tau)/p_0)^n} - \mu_m \mu \quad \frac{dp}{dt} = \mu - \mu_p p$$

where μ_m and μ_p are decay rates, p_0 is the repression threshold, n is a Hill coefficient and τ is the time delay. The application of our method to DDEs is of particular interest since unobserved species or components of a system can often be modelled in terms of a time delay. In addition, numerical solutions to DDEs are generally much more computationally expensive to obtain than ODEs, and thus inference of such models using MCMC methods with the more direct approach of explicitly solving the system at each iteration becomes less feasible as the complexity of the system of DDEs increases.

5.2 Gaussian Processes for Approximate ODE Inference

Monk DDE Model					
Samples	Method	μ_m	$\mu_p \times 10^{-3}$	$p_0 \times 10^{-3}$	τ
40	GP MAP	100.21 ± 2.08	29.7 ± 1.6	30.1 ± 0.3	25.65 ± 1.04
	GP Full Bayes	99.75 ± 1.50	29.8 ± 1.2	30.1 ± 0.2	25.33 ± 0.85
80	GP MAP	99.48 ± 1.29	29.5 ± 0.9	30.1 ± 0.1	24.81 ± 0.59
	GP Full Bayes	100.26 ± 1.03	30.1 ± 0.6	30.1 ± 0.1	24.87 ± 0.44
120	GP MAP	99.91 ± 1.02	30.0 ± 0.5	30.0 ± 0.1	24.97 ± 0.38
	GP Full Bayes	100.23 ± 0.92	30.0 ± 0.4	30.0 ± 0.1	25.03 ± 0.25

Table 5.3: Summary statistics for each of the inferred parameters of the Monk model

We consider data generated from the above model, with parameters $\mu_m = 0.03$, $\mu_p = 0.03$, $p_0 = 100$, $\tau = 25$, at $\{40, 80, 120\}$ time points with added random noise drawn from a Gaussian distribution, $N(0, v)$ for $v = 0.1 \times \sigma_n$, where σ_n is the standard deviation of the time-series data for the n th species. The parameters were then inferred from these data sets using our GP-based population MCMC methods. Figure 5.4 shows a time comparison for 10 iterations of the GP sampling algorithms and compares it to explicitly solving the DDEs using the Matlab solver DDE23 (which is generally faster than the Sundials solver for DDEs). The GP methods are around 400 times faster for 40 data points. Using the GP methods, samples from the full posterior can be obtained in less than an hour. Solving the DDEs explicitly, the population MCMC algorithm would take in excess of two weeks computation time, assuming the chains take a similar number of iterations to converge.

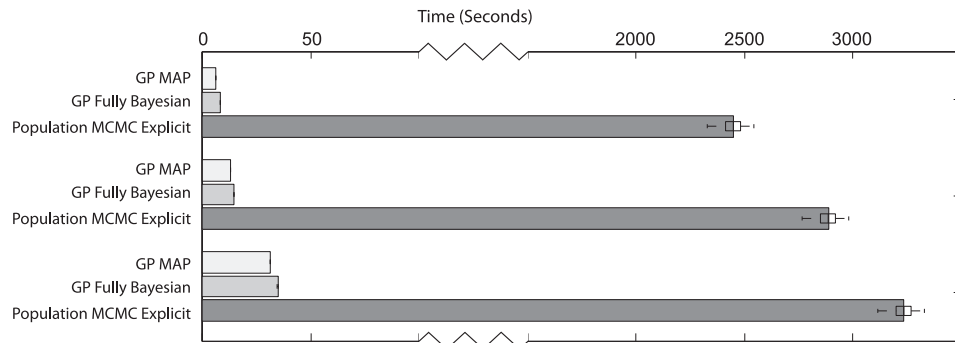


Figure 5.4: Summary statistics for time taken to perform DDE inference using a variety of methods - Summary statistics of the time taken for the algorithms to complete 10 iterations using DDE model.

5.2.8 Example 3 - The p53 Gene Regulatory Network with Unobserved Species

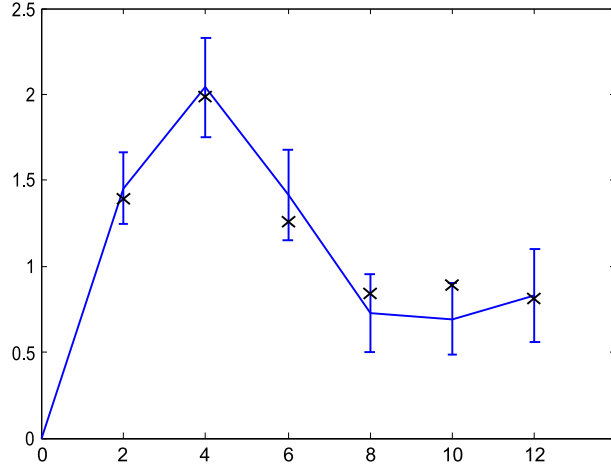


Figure 5.5: The predicted output of the p53 gene using data from (12) and the accelerated GP inference method for the linear model - The predicted output of the p53 gene using data from (12) and the accelerated GP inference method for the linear model. Our results are compared to the results obtained in (12) (shown as crosses) and are comparable to those obtained in (69).

Our third example considers a linear and a nonlinear model describing the regulation of 5 target genes by the tumour repressor transcription factor protein p53. We consider the following differential equations which relate the expression level $x_j(t)$ of the j th gene at time t to the concentration of the transcription factor protein $f(t)$ which regulates it, $\dot{x}_j = B_j + S_j g(f(t)) - D_j x_j(t)$, where B_j is the basal rate of gene j , S_j is the sensitivity of gene j to the transcription factor and D_j is the decay rate of the mRNA. Letting $g(f(t)) = f(t)$ gives us the linear model originally investigated in (12), and letting $g(f(t)) = \exp(f(t))$ gives us the nonlinear model investigated in (69). The transcription factor $f(t)$ is unobserved and must be inferred along with the other structural parameters B_j , S_j and D_j using the sampling scheme detailed earlier in this chapter. In this experiment, priors on the unobserved species used were $f(t) \sim \Gamma(2, 1)$ with a log-Normal proposal. We tested our method using the leukaemia data set studied in (12), which comprises 3 measurements at each of 7 time points for each of the 5 genes. Figures 5.5 and 5.6 show the inferred missing species and the results are in good accordance with recent biological studies. For this example, our GP sampling algorithms

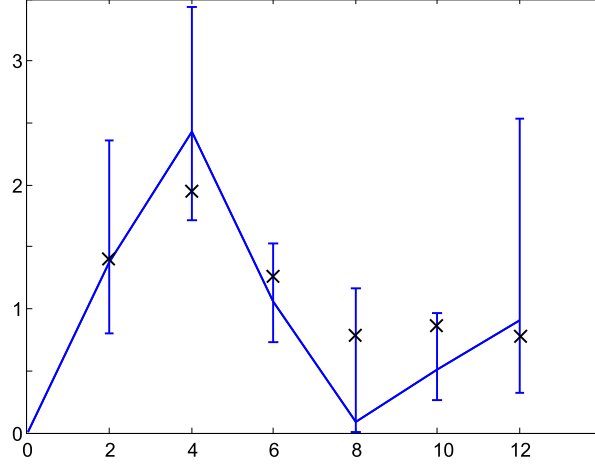


Figure 5.6: The predicted output of the p53 gene using data from (12) and the accelerated GP inference method for the nonlinear model - The predicted output of the p53 gene using data from (12) and the accelerated GP inference method for the nonlinear model. Note that the asymmetric error bars in the nonlinear model are due to $\exp(\mathbf{y})$ being plotted, as opposed to just \mathbf{y} for the linear model. Our results are compared to the results obtained in (12) (shown as crosses) and are comparable to those obtained in (69).

ran to completion in under an hour on a 2.2GHz Centrino laptop, with no difference in speed between using the linear and nonlinear models; indeed the equations describing this biological system could be made more complex with little additional computational cost.

5.2.9 Discussion

Calculating explicit solutions to differential equations can pose computational challenges for the application of inferential methodology. An alternative approach to this problem has been suggested in previous work based on employing computationally inexpensive spline-based approximations to provide surrogate solutions to the ODE system. Such approaches are fast, although generally prone to errors that propagate from the spline approximation step; overfitting the data is a particular problem.

In this section we have extended this approach by framing it within a Bayesian framework and employing an auxiliary Gaussian process model, which allows uncertainty in the data smoothing step to be better characterised. The hyperparameters that control the level of smoothing may be inferred and this automatically provides a

parsimonious fit to the data. There are a couple of drawbacks to this approach however. There must be enough data for the smoothing step to give reasonably accurate derivative estimates, and it is well known that GPs scale badly with the number of observations. In addition the chosen covariance function must be flexible enough to describe the dynamics of the data. Further it is not possible to estimate marginal likelihoods, since the likelihood is based on a smoothing of the data via a GP, instead of being based on the data directly. Consequently, although this GP based approach offers great speed advantages for parameter inference in certain cases, we must conclude that it is not so useful for model inference when we are most interested in the underlying structure of the system and wish to estimate marginal likelihoods for model ranking. For the rest of this thesis we therefore focus on manifold sampling methods that require explicit solution of the systems of differential equations.

5.3 Disease Outbreak Models

One of the main aims in this thesis is to be able to tackle the motivating example of modelling realistic networks of biochemical dynamics. In this chapter we have so far investigated methodology that may be applied to models based on differential equations. Before we tackle larger and more complex systems, such as the circadian rhythm model we consider in Chapter 6, let us first investigate some of the modelling issues associated with these types of mechanistic models. In this section, we shall examine some simple ODE models describing the outbreak of a disease, which allows us to more closely examine in a pedagogical manner the factors that may affect the results of a Bayesian parameter and model analysis. The individual states within a complex system can often be naturally described in terms of their rates of change, which can be defined in terms of the interactions that take place. As a result, ordinary differential equations (ODEs) have proved to be extremely useful tools for describing, in quantitative terms, how physical systems change over time. Accurately modelling complex systems using systems of ODEs that have some explanatory and predictive power involves identifying the essential features of the system and representing these mathematically.

In the case of disease outbreaks, just as in any other modelling context, a good model therefore allows us to

explain the main interactions present within a system and the key mechanisms at work,

predict the future dynamics of the system,

investigate hypothetical scenarios that are not amenable to experimental study.

Such models are of particular use in ecology, epidemiology and biology, where there are many unobservable states and direct experimentation with the system of interest may be very difficult or even impossible. The main challenge with these ODE models is that they involve parameters whose values are also generally not observable. These may correspond to rates of infection or interaction, which are often impossible to measure experimentally. Such parameter values must therefore be inferred from the available observed data, ideally using a probabilistic approach that can give a measure of uncertainty in the answers.

We can use these models to answer specific questions, for example, if the number of observed people with a particular disease in a town over the past 5 days is 123, 127, 104, 92, 74, what is the likelihood of the infection spreading over the coming weeks and when should I try to leave in order to minimise risk of infection? Using some simple models motivated by (142), we will consider how much prior knowledge and how much observational data is needed in order to make useful inferences about the spread of a disease. We will also consider the issue of comparing model hypotheses. Suppose we have two theories regarding how the disease spreads

- (a) infection occurs from contact with one other infected person
- (b) infection occurs from contact with two other infected people

We can encode both of these hypotheses in the form of ODE models, and employ a Bayesian analysis to investigate which better explains the observed data; this kind of Bayesian approach to model ranking is a topic that has only recently been tackled systematically (27, 203).

5.3.1 A Simple Infection Model

We now illustrate some key ideas by focusing on a very simple disease outbreak model. We will let $S(t)$ and $I(t)$ represent the number of healthy people and diseased people at time t . Since $S(t)$ and $I(t)$ will take real values, it is more appropriate to think of these states as representing the concentration levels of the two species. We suppose that there is just one event that can cause a change in these concentrations: a healthy person may come into contact with a diseased person and contract the infection with some probability. We note immediately that in this simple world the population of healthy people will decrease to zero and the population level of diseased people will correspondingly increase until everyone has become infected. We may model this assumption by introducing a single parameter, a rate constant, β , that characterises the rate of infection. The larger the value of β , the more virulent the disease. Simple mass action modelling then leads us to the ODEs

$$\dot{S}(t) = -\beta S(t)I(t), \quad (5.28)$$

$$\dot{I}(t) = \beta S(t)I(t). \quad (5.29)$$

So the ODE system (Equations 5.28 - 5.29) specifies the rate of change of the two population levels. Adding the two equations, we see that the rate of change of the total population is zero, which assumes that this particular disease is not fatal, at least not for the modelling time period we consider. This is intuitively obvious, since each time we lose a healthy person we gain a corresponding infected person, and so $S(t) + I(t)$ remains constant. Letting the constant K denote this overall population size, we have $S(t) + I(t) = K$. We may then replace $I(t)$ by $K - S(t)$ in Equation 5.28 to get a single ODE

$$\dot{S}(t) = -\beta S(t) (K - S(t)). \quad (5.30)$$

This ODE fits into the class of logistic equations. Although nonlinear, it is sufficiently simple to admit a pencil-and-paper solution. This may be written

$$S(t) = \frac{S(0)K}{S(0) + (K - S(0))e^{\beta Kt}}. \quad (5.31)$$

Here $t = 0$ represents the initial time when we start to monitor the populations, so $S(0)$ is the initial level of the human population.

Let us first assume that the size of the total population, K , is known, and that the initial number of healthy people, $S(0)$, is also known. This leaves us with one unknown parameter, β , which may be interpreted as the rate at which healthy people contract the disease and is measured per infected person per day. Therefore if initially there are 1000 healthy people and just 1 infected person, a rate of $\beta = 0.001$ initially corresponds to $S(0) \times \beta = 1000 \times 0.001 = 1$ person being infected by each diseased person per day. We note that the number of people becoming infected depends on both the number of healthy people and the number of currently infected people at any particular time; by solving our differential equations we may see how these two populations evolve relative to one another.

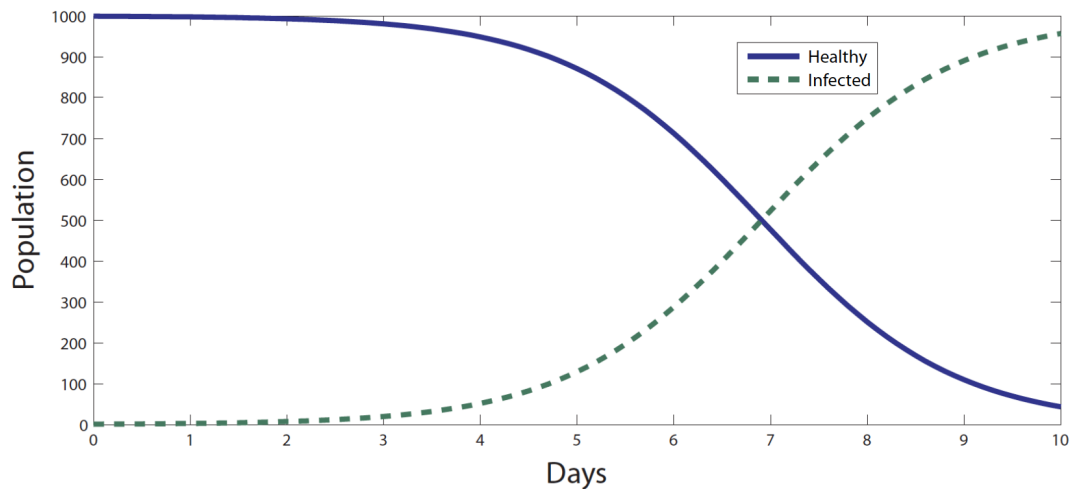


Figure 5.7: Example output from a simple infection model - Sample solution of the simple infection ODE with parameter $\beta = 0.001$.

Figure 5.7 shows how a population of 1000 healthy individuals diminishes to less than 100 after just 10 days, based on a single infected person within the population and the infection spreading at a rate of one person per infected person per day. We may see the effect of doubling the rate constant to $\beta = 0.002$ in Figure 5.8. In this case the population dwindles to less than 100 after just 5 days, and by the 7th day everyone has effectively been infected.

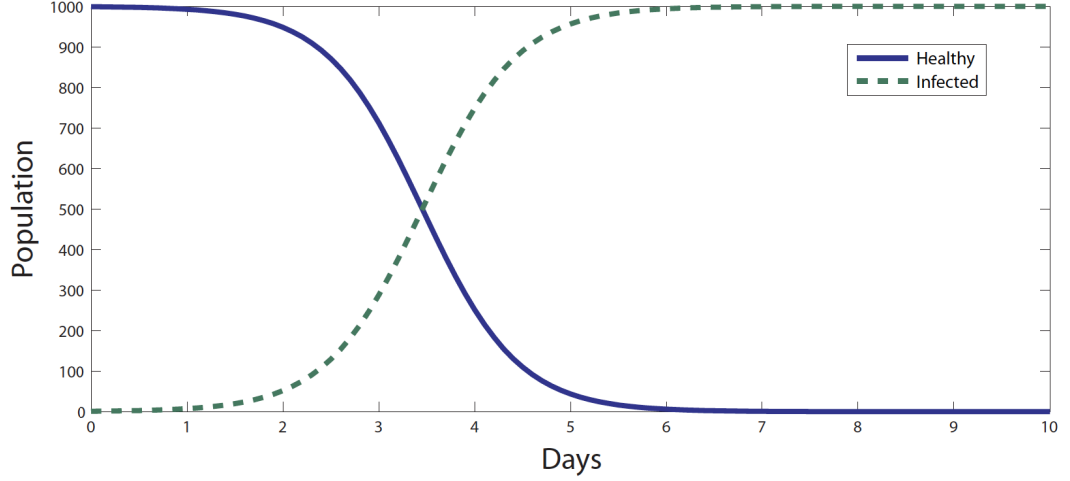


Figure 5.8: Example output from a simple infection model - Sample solution of the simple infection ODE model with parameter $\beta = 0.002$.

These predictions have been made under the assumption that we know the exact infection rate. Given an initial number of healthy and infected individuals, along with a rate constant, our model tells us how the population evolves. In a more realistic scenario, we would not know this rate constant, but instead have (usually inaccurate) observations of how the population changes over a period of time. The goal in this case is then reversed; given some observations regarding the number of infected people at certain time points, we want to estimate the rate constant β such that our model best describes the situation as we see it, and we do this by employing Bayesian methods. Once we have inferred the rate parameter β , we may then quantify the probability of future scenarios, based on parameter values such that our model adequately describes the past.

There are three levels of inference that we may ultimately wish to carry out. The first determines the parameters with which the model plausibly describes the data. This is the probability of the free parameters $\theta = [\theta_1, \dots, \theta_n]$ given some data \mathbf{Y} and a particular model M , which may be written as $P(\theta|\mathbf{Y}, M)$. In our example there is just one parameter, $\theta = \beta$ and \mathbf{Y} is a vector of observations at a number of time points. The second level of inference sheds light on the uncertainty associated with our choice of model, and this is the probability of a particular model M given the data \mathbf{Y} , written as $P(M|\mathbf{Y})$. Finally, the third level of inference describes the probability of a *prediction*

given the data, and this prediction may be based on multiple plausible models which are weighted according to their relative probabilities.

Let us first define our prior and likelihood distributions for the model describing simple infection dynamics. A prior can be constructed by considering a reasonable time scale for the process; we may argue that not everyone will become infected in less than a single day. This gives an upper limit for the rate constant of $\beta = 1$, since with that value $S(0) \times \beta = S(0)$, so all $S(0)$ people could be infected by one diseased person on the very first day. Likewise, a reasonable minimum rate is $\beta = 0$, in which case nobody else contracts the disease. With no further information, it is reasonable to take the view that, a priori, any rate constant between 0 and 1 is equally likely. So we may choose our prior on β to have a uniform distribution over this range.

The choice of probability distribution to describe the presumed measurement error depends on the problem context. For some modelling scenarios, where for example the observed data is the number of counts occurring within a particular time interval, the choice of a Poisson distribution may be appropriate. Alternatively, if the observed data is obtained from estimates that may be affected by a large number of small unknown random factors, then due to the Central Limit Theorem (97, 180) the associated error may be well approximated by a Gaussian distribution. In the case of modelling our disease outbreak, we shall assume that the estimated population levels are subject to small unknown errors that affect the count of infected people. We also assume that the errors at different observation times are independent, and therefore define the likelihood function to be a product of Gaussian distributions,

$$L = P(\mathbf{Y}|\boldsymbol{\theta}, M) = \prod_t N_{Y(t)}(S(t), \sigma^2). \quad (5.32)$$

The variance σ^2 can either be estimated and fixed in advance, or inferred along with the other parameters. In the case of our simple infection model, $S(t)$ is easily calculated from Equation 5.31, however for more complex models it is necessary to compute a numerical solution for the ODE model. Noting that the marginal likelihood $P(\mathbf{Y}|M)$ is constant for a particular model M , we see it may be calculated as the integral of the likelihood times the prior over all parameter values,

$$\begin{aligned}
 P(\mathbf{Y}|M) &= \int \dots \int P(\mathbf{Y}, \boldsymbol{\theta}|M) d\theta_1 \dots \theta_n \\
 &= \int P(\mathbf{Y}|\boldsymbol{\theta}, M) P(\boldsymbol{\theta}|M) d\boldsymbol{\theta}.
 \end{aligned}$$

Performing the second and third levels of statistical inference over ODE models is challenging precisely because we have to estimate such an integral, which is generally analytically intractable and high dimensional. Only recently has it been demonstrated that this integral may be efficiently and accurately estimated using thermodynamic integration (27), and we shall employ this later to discriminate between competing model hypotheses describing disease outbreaks and to predict future infection dynamics.

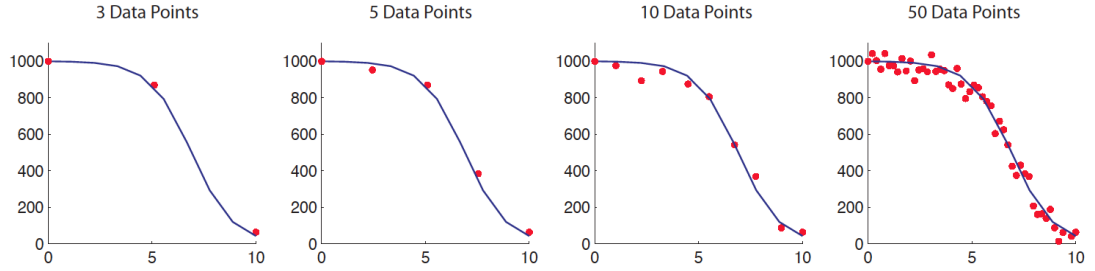


Figure 5.9: Varying numbers of data points generated from a simple infection ODE model - [3, 5, 10, 50] data points generated from the simple infection model over 10 days with parameter $\beta = 0.001$ and Gaussian distributed noise with a standard deviation of 50.

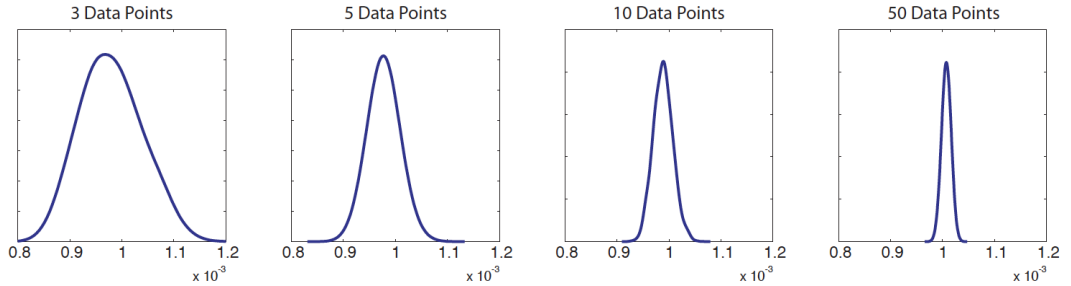


Figure 5.10: Posteriors inferred from a simple infection ODE model with a varying number of data points - Posterior output from the simple infection model with parameter $\beta = 0.001$ and Gaussian distributed noise with a standard deviation of 50. As the number of data points increases, so the posterior becomes more sharply peaked around the true value of β .

Using the simplified mMALA approach described in Section 5.1, we now infer the posterior distribution over the parameter values given some synthetic data, which allows us to judge the performance of the algorithm under controlled conditions. We evaluate the solution of the differential equation 5.31 for a chosen value of β at a number of time points and add Gaussian distributed noise with known variance to the solution to generate some experimental data. We generated four data sets this way, as shown in Figure 5.9. We can then treat this data as though it came from the model with an unknown value of β , and investigate the effect this has on the variance of the inferred posterior distribution. Figure 5.10 shows how the posterior distribution over β becomes more sharply peaked around the true value as the number of data points increases from 3 to 50. The noise induces a noticeable bias when using just 3 data points, although the posterior probability is reasonably large at the true value of 0.001. Figures 5.11 and 5.12 show how the posterior distribution becomes less diffuse as we add less noise, indicating a greater confidence in the range of values for which the model could plausibly describe the data. We may also examine the effect of our prior on the posterior. Changing the prior from uniform over $[0, 1]$ to uniform over $[0, 0.01]$ has very little difference on the posterior of β , as shown in the left and middle pictures of Figure 5.13. However, if we were to badly mis-specify the prior, for example by setting it to be a low variance Gaussian distribution over the wrong value, we observe the rather biased posterior distribution shown in the picture on the right in Figure 5.13. This type of mis-specified prior may be diagnosed by comparing the overlap of the prior and posterior.

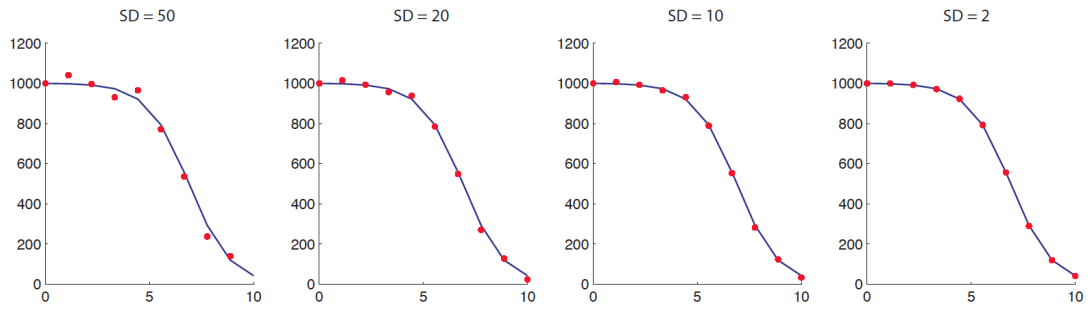


Figure 5.11: Data points generated from a simple infection ODE model with varying noise - 10 data points generated from the simple infection model over 10 days with parameter $\beta = 0.001$ and Gaussian distributed noise with a standard deviations $[2, 10, 20, 50]$.

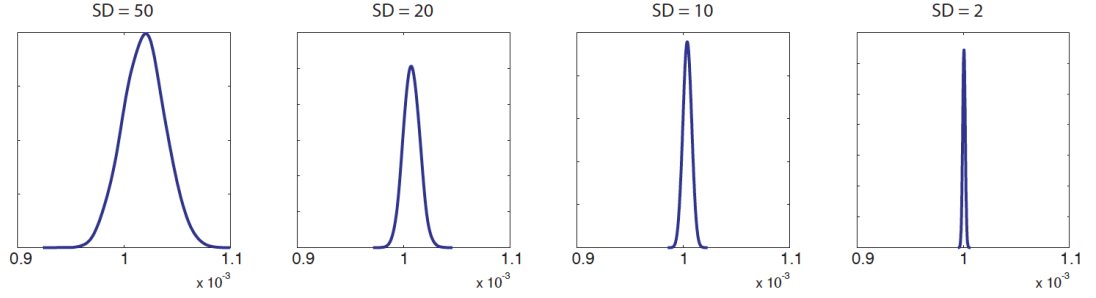


Figure 5.12: Posterior distributions inferred from simple infection ODE model with varying noise - Posterior output from the simple infection model with parameter $\beta = 0.001$ and 10 data points. As the standard deviation of the additive noise decreases, so the posterior becomes more sharply peaked around the true value of β .

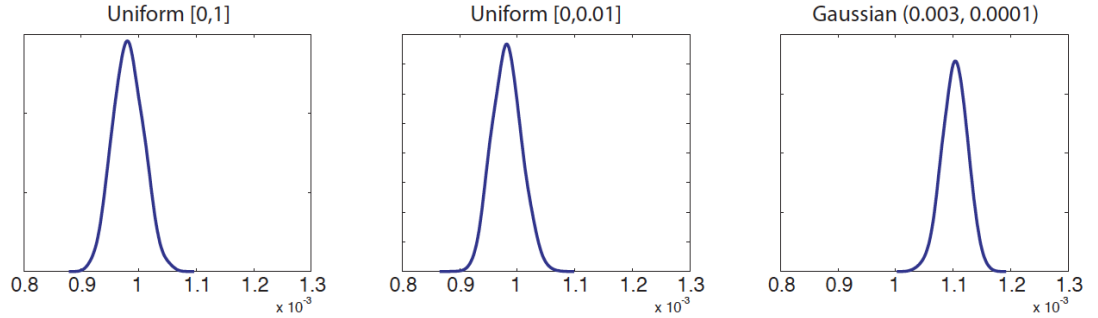


Figure 5.13: Examining the effect of the prior on posterior inference over a simple infection ODE model - Posterior output from the simple infection model with parameter $\beta = 0.001$, 3 data points and Gaussian noise with standard deviation 20. Changing the uniform prior from the range $[0, 1]$ to $[0, 0.01]$ has little effect on the posterior. However, a mis-specified Gaussian prior with mean 0.003 and standard deviation 0.0001 cannot be properly corrected by this small amount of data.

We have investigated the effects of data and priors when performing a Bayesian analysis of a very simple ODE model describing spread of infection. In the next section we will consider inference over slightly more complex models with partial observations.

5.3.2 A More Realistic Infection Model

The model derived in Section 2 of (142) splits the overall population into three classes. At each time t we have

- Healthy individuals, $S(t)$,
- Infected individuals, $I(t)$,
- Individuals with dormant infection, $R(t)$.

As in our simple model (Equations 5.28–5.29), healthy individuals may be infected by coming into close contact with infected individuals. However, we now allow for the possibility that the disease becomes dormant in an infected individual. We make the assumption that people with the dormant infection are unable to be counted as they may display no symptoms, and that their infection may become active again with some probability. These modelling assumptions lead us to the following ODE system adapted from (142), and for simplicity we consider only the short timescale regime ($\Pi = \delta = 0$ in (142)), such that the total population size is constant.

$$\dot{S}(t) = -\beta SI, \tag{5.33}$$

$$\dot{I}(t) = \beta SI + \zeta R - \alpha SI, \tag{5.34}$$

$$\dot{R}(t) = \alpha SI - \zeta R. \tag{5.35}$$

The parameters in this model are:

- β : an infection rate constant similar to that in the simple model of Section 5.3.1. A larger value of β implies a more virulent infection.
- α : a rate constant relating to the probability that the disease in an infected person becomes dormant. A larger value of α implies a higher probability of dormancy.

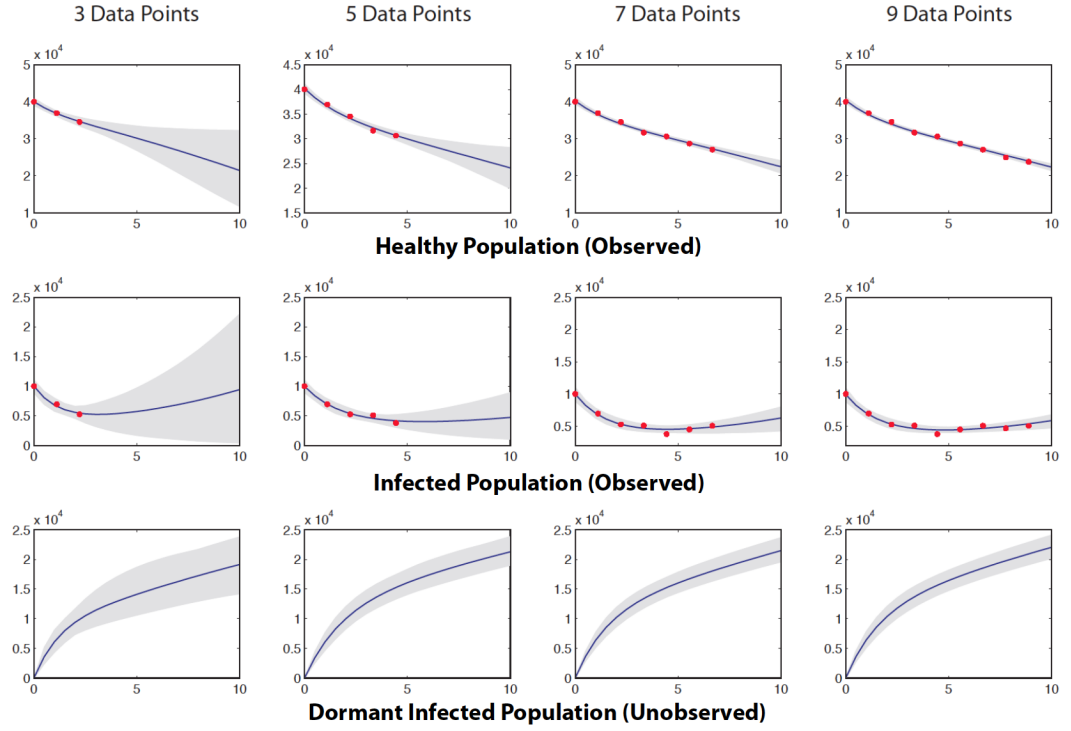


Figure 5.14: Examining the effect of the number of data points on posterior output from the complex infection ODE model - Posterior output from the complex infection model (Equations 5.33–5.35) with parameters $\beta = 0.00001$, $\alpha = 0.00002$, $\zeta = 0.1$, and Gaussian distributed noise with a standard deviation of 500. As the number the of data points increases, so the uncertainty in the posterior model output decreases.

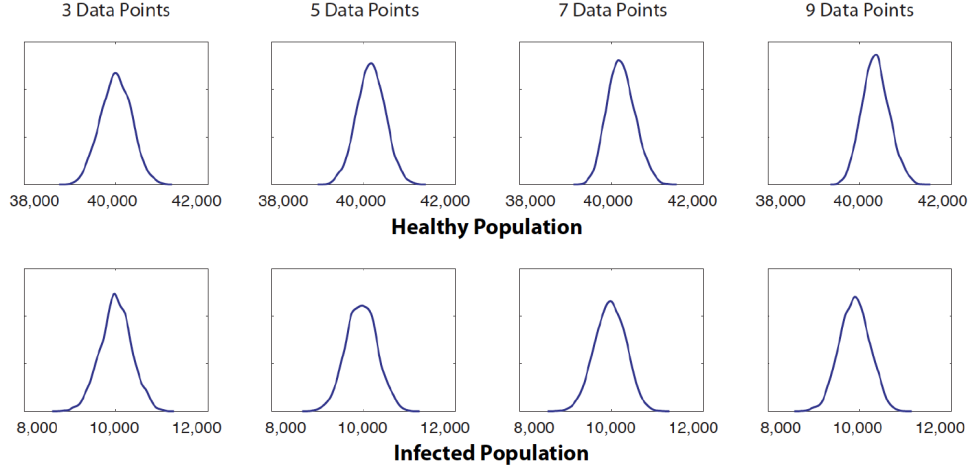


Figure 5.15: Posterior inference over the initial conditions of a complex infection ODE model - Posterior distributions of the inferred initial conditions from data generated by the complex infection model with parameters $\beta = 0.00001$, $\alpha = 0.00002$, $\zeta = 0.1$, and Gaussian distributed noise with a standard deviation of 500. We observe that the initial conditions are in this case relatively insensitive to the number of data points observed.

- ζ : a rate constant for a dormant infection to become active again. A larger value of ζ implies a higher probability of the disease becoming active.

Here we consider a disease outbreak in a mid-sized town. As in Section 5.3.1, we generate artificial data from the ODE model and investigate the variance of the inferred posterior distributions. We infer the initial conditions for the infected and healthy individuals, as well as all parameter values. We assume that initially there are no individuals with a dormant infection, and that there are around 40,000 healthy people and already 10,000 infected people living in the town. Given daily observations over a period of $[3, 5, 7, 9]$ days, we consider the predictive model output, that is, two standard errors or a 95% confidence for the output of the ODE at each time point, over the 10 days after the disease is first observed, shown in Figure 5.14. Figure 5.15 shows also the inferred initial conditions, which in this example are relatively insensitive to the number of observed data points. As the number of observed data points increases, however, we see that the uncertainty in the predictive model output over subsequent days does indeed decrease. This is also seen clearly when we consider the predictive posterior model output for day 15, given observations over $[3, 5, 7, 9]$ days, as shown in

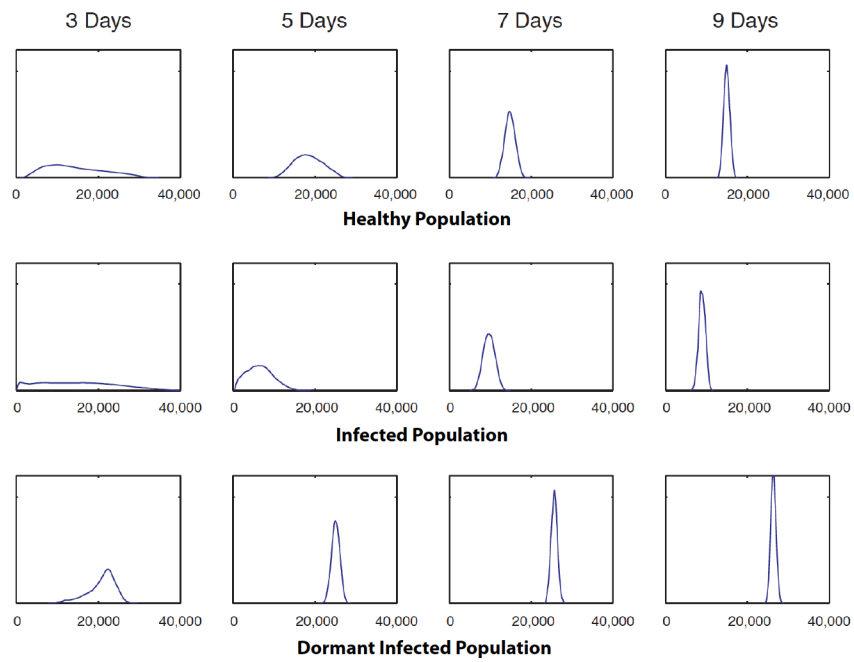


Figure 5.16: Prediction of future infection levels from a complex infection ODE model - Predicted levels of healthy people, infected people and people with dormant infection on day 15 having observed levels of healthy individuals and infected individuals for $[3, 5, 7, 9]$ days.

Figure 5.16. After just 3 days of observations, we can say very little about the predicted healthy and infected population sizes on day 15. Given an additional observation on each of the following two days however, we can predict with much greater certainty that the number of people who escape infection is likely to be between 10,000 and 25,000. As we collect more data over the subsequent days, our predictions become more and more confident, as seen from the more sharply peaked posterior distributions over the each population. Indeed, after 7 and 9 days, the predicted number of infection-free individuals is roughly between 10,000 and 18,000, and 13,000 and 17,000 respectively. These predicted ranges are tending towards the “true” number of healthy individuals, determined by the system of ODEs to be 14,790. Likewise the predicted numbers of people with active and dormant infection tend towards their “true” values of 10,426 and 24,784, respectively.

5.3.3 Model Selection

We now consider a second level of inference, in which there is uncertainty not only in the parameters, but also in the specified model. Suppose we make the assumption that an individual can only become infected when they come into contact with two people who have an active infection. Applying similar arguments to those used for Equations 5.33 to 5.35 we may arrive at the alternative set of ODEs

$$\dot{S} = -\beta SI^2, \tag{5.36}$$

$$\dot{I} = \beta SI^2 + \zeta R - \alpha SI, \tag{5.37}$$

$$\dot{R} = \alpha SI - \zeta R. \tag{5.38}$$

We refer to Equations 5.33 to 5.35 as Model 1, and Equations 5.36 to 5.38 as Model 2.

Given some data, we may now perform parameter inference over each model. Nine days observed data was generated by simulating from Model 1, in which infection can spread from a single individual, and adding some Gaussian distributed noise with standard deviation 500. We wish to determine whether a Bayesian inference will allow us to conclude that Model 1 describes this data better than Model 2, assuming we do not have any detailed information about the rate constants. Figures 5.17, 5.18 and 5.19 show the posterior model outputs for the two proposed models, setting the standard

deviation of the noise in the model to be [500, 1000, 2000] respectively. We note that visually trying to assess which is the better model is difficult, since the posterior output covers most of the data points for both models.

We may therefore resort to calculating Bayes factors such that B_{12} represents the weight of statistical evidence in favour of Model 1 over Model 2. This is computed as the ratio of the marginal likelihoods for the two competing models,

$$B_{12} = \frac{P(\mathbf{Y}|M_1)}{P(\mathbf{Y}|M_2)}. \quad (5.39)$$

We recall that calculating the marginal likelihood involves estimating the integral of the posterior over all values of the parameters, which is a rather challenging task. We employ the technique of thermodynamic integration, which has recently been shown to provide accurate, low variance estimates of this quantity (67), whereas other seemingly simpler methods, such as the Posterior Harmonic Mean estimator, may fail to produce usable results (27). We make use of the population MCMC approach described in Section 3.5, again using the simplified mMALA algorithm from Section 5.1 to draw samples from each of the tempered distributions.

Table 5.4 is a useful guide for interpreting the evidence provided by the estimated Bayes factors (102). Table 5.5 shows the results of the marginal likelihoods estimated 10 times for each model. In each case, the Bayes factors correctly identify the model that was used to produce the data. As the standard deviation of the additive noise increases from 500 to 1000 to 2000, the weight of evidence as indicated by the Bayes factors decreases, although the evidence remains substantially in favour of the correct model.

Table 5.4: Interpretation of Bayes Factor

B_{12}	Evidence against H_2
1 to 3	Not worth more than a bare mention
3 to 10	Substantial
10 to 100	Strong
> 100	Decisive

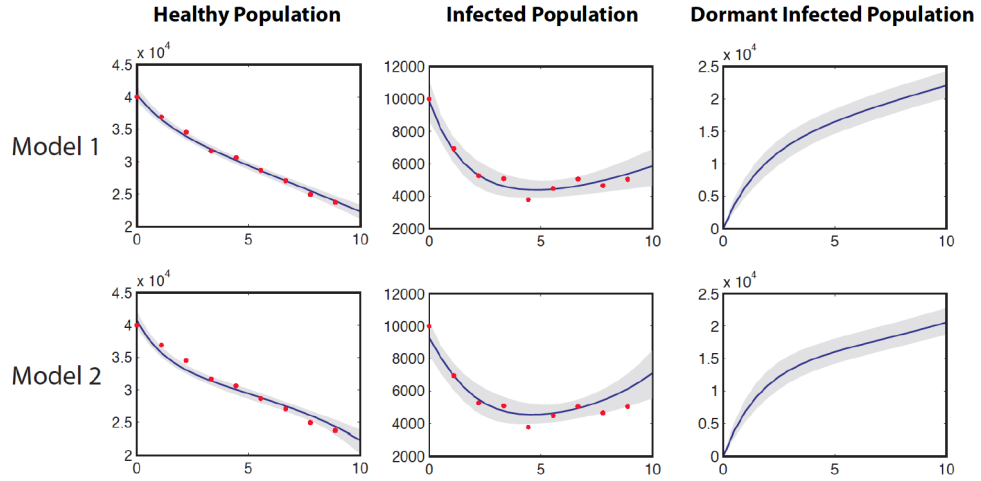


Figure 5.17: Comparison of posterior outputs from two plausible infection models with low noise levels - Posterior output for the two competing model hypotheses with the standard deviation of noise set to 500.

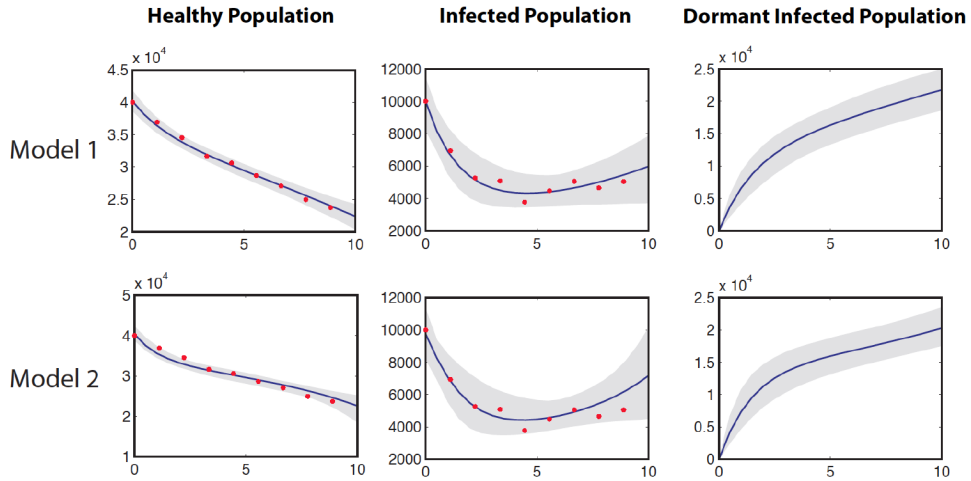


Figure 5.18: Comparison of posterior outputs from two plausible infection models with medium noise levels - Posterior output for the two competing model hypotheses with the standard deviation of noise set to 1000.

Table 5.5: Summary of estimated marginal likelihoods for each infection model

Model	Noise SD	Marginal Likelihood (\pm Standard Error)	Bayes Factor B_{12}
Model 1	500	-152.7 (\pm 0.1)	31.3 (\pm 1.7)
Model 2	500	-184.0 (\pm 1.6)	
Model 1	1000	-158.5 (\pm 0.1)	16.7 (\pm 1.0)
Model 2	1000	-175.2 (\pm 0.9)	
Model 1	2000	-167.1 (\pm 0.1)	9.8 (\pm 3.3)
Model 2	2000	-176.9 (\pm 3.2)	

5.3.4 The Optimal Time to Escape

Having introduced the Bayesian approach for performing inference over systems of differential equations to describe disease outbreaks, we now return to the scenario posed in the introduction of this section. The number of observed people with a particular disease in a town over the past 5 days is [123, 127, 104, 92, 74]. What is the likelihood of the infection spreading over the coming weeks and when should I try to leave in order to minimise risk of infection?

We now know how to address this question through the use of Bayesian modelling. We shall assume Model 1 to be a fair representation of the interaction between the infected population, healthy population and the dormant population, noting that if we had multiple plausible models we could of course once again do full model comparison by calculating Bayes factors. We perform parameter inference over the model given our data of reported daily numbers of diseased people and plot the predictive model output, shown on the left of Figure 5.20.

In this case, the 95% confidence output from the model includes a wide variety of outcomes, and the uncertainty in the estimate naturally increases with time. We cannot rule out the scenarios where (a) there is relatively little spread of disease within the population, or (b) almost everyone becomes infected within the next month. The mean number of infected individuals continues to decrease over the next 5 days, until day 10 when it begins to increase. We therefore argue in favour of making an early exit in order to minimise risk of exposure to an actively infected individual, since after day 10 there is much more uncertainty regarding population estimates.

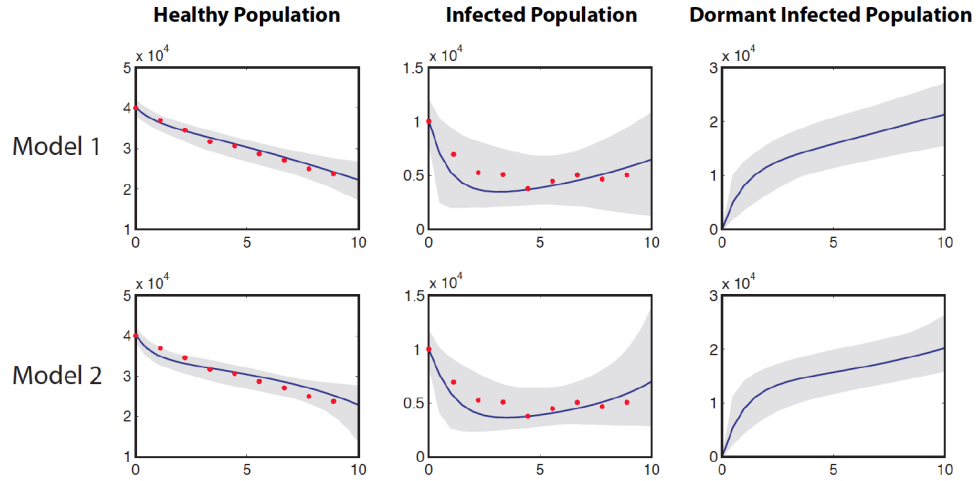


Figure 5.19: Comparison of posterior outputs from two plausible infection models with high noise levels - Posterior output for the two competing model hypotheses with the standard deviation of noise set to 2000.

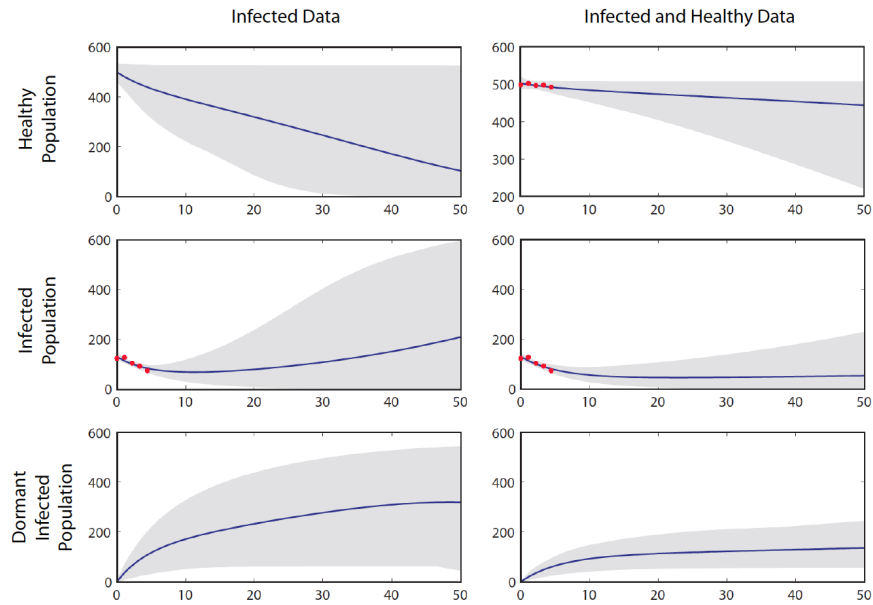


Figure 5.20: Model predictions of disease outbreak in a small town - The left hand side plots show the predictive model output from the 3 parameter model given observations of infections in a small town. The right hand side plots show the predictive model output given additional observations of the number of healthy individuals in the population.

If we have additional data regarding the likely number of healthy individuals in the town over the past 5 days, then we may again perform parameter inference over our model, this time including data for both healthy and infected individuals. The additional data and predictive model output is shown on the right hand side of Figure 5.20. Given this extra data, we can say with much more certainty that the number of actively infected individuals in the town will remain low for a longer period of time, making it perhaps less urgent for us to make our getaway immediately.

5.3.5 Discussion

Mathematical modelling of natural phenomena has a long and illustrious history. Differential equations have the potential to describe and predict the behaviour of many different types of complex systems. However, any mathematical model represents an abstracted summary that cannot be claimed to capture all characteristics of interest. A well-known quote that is paraphrased from (21) and attributed to George Box says that “All models are wrong but some are useful”. Modelling involves compromises and it generates an inherent level of uncertainty. Moreover, the task of identifying unknown or unmeasured model parameters introduces further uncertainty.

In cases where model predictions may be used to guide policy, for example, in economics, weather forecasting and epidemiology, a systematic and consistent treatment of all levels of model uncertainty is vital. We have demonstrated in this section, using simple examples, that the Bayesian statistical framework provides an appropriate means with which to capture information and reason under uncertainty, and we have shown how this Bayesian inferential methodology may be extremely useful, not only for considering individual statistical models based on systems of differential equations but also for systematically comparing different model hypotheses given limited and uncertain observed data.

5.4 Conclusions

In this chapter we have seen two different Bayesian approaches for parameter inference over models described by nonlinear differential equations. The most obvious computational bottleneck when inferring parameters over such models is the required numerical solution of the system of differential equations. We have seen how data smoothing may

be used to approximate the data and provide derivative estimates, such that we can perform inference in the derivative space, instead of the state space, and the method we have presented offers a number of advantages. The use of a Gaussian process instead of a smoothing spline can help prevent overfitting to the data. Parameter inference can be faster using the surrogate likelihood based on the derivative mismatch, instead of explicitly solving the system of ODEs at each iteration. It has been noted that posing the problem in the derivative space has a smoothing effect on the posterior distribution, making it easier to optimise or sample from (159), and with the Bayesian approach described in this chapter, we may also implement our sampling scheme within a population MCMC framework, which allows the sampler to escape from potential local maxima induced by nonlinearities in the ODEs.

However, the main problem is that marginal likelihoods cannot be calculated using this GP approach, since inference is based on the approximate smoothing of the data instead of the data directly. With sparse measurements there may be large amounts of uncertainty in the Gaussian process approximation, resulting in poor derivative estimates. Such an approach seems perhaps most useful for parameter inference, particularly for delay differential equations, which can be extremely computationally expensive to solve. Ultimately however, the inability to calculate marginal likelihoods prevents us from performing full model comparison over hypothesised structures.

We therefore advocate the use of the explicit method, developed in Section 5.1, for model ranking purposes, whereby we may obtain more efficient sampling proposals by exploiting the local geometry of the parameter space. Explicitly solving the system of differential equations at each iteration can be costly, however it allows the full Bayesian machinery to be used, in particular the ability to estimate marginal likelihoods for model comparison. By using the local sensitivity information at each point, moves can be proposed with greater efficiency, and we saw that in particular the simplified mMALA method offered the best performance of any of the MCMC methods tested on these statistical models. Finally we note that the computational expense of obtaining explicit solutions can be remedied somewhat by parallelising the population MCMC code, which makes application to larger ODE models much more feasible, as we shall see in Chapter 6.

6

Modelling Biochemical Dynamics

We now consider the challenge of mathematically modelling biochemical dynamics, and return to the motivating example of circadian rhythms in plants, employing the differential geometric sampling methodology developed in the previous chapters to investigate and rank statistical models based on systems of nonlinear differential equations to describe such rhythmic behaviour. Interestingly it was once again Darwin who first suggested the heritability of circadian rhythms in plants (48, 132), implying that although plants could adapt to their environment their natural behaviour was mainly defined by the underlying biochemical structure. Mathematical modelling of the biochemical processes of plants, and indeed other organisms, allows deeper insight to be gained into the inner workings of a variety of physiological mechanisms originating at the molecular level. In this chapter, we also consider inference over an ODE model of a cell signalling pathway that is based on mass-action kinetics and exhibits a transient response, in contrast to the limit-cycle behaviour of the circadian model. This work follows from an invited paper published in the Journal of the Royal Society, Interface Focus (28).

6.1 Mathematical Modelling

The use of mathematical modelling has long played an important role in describing and predicting the behaviour of natural processes (73, 110, 195). It is only more recently however that the use of such models within a statistical framework has allowed modelling and experimental approaches to be more tightly bound together forming an

iterative, more symbiotic relationship (139, 207). We recall that mathematical models are abstract representations of reality that are useful for making testable predictions with varying levels of detail. Deterministic nonlinear ordinary differential equations (ODEs) for example may be most appropriate when describing the average concentration of a protein within a population of cells, in which the stochastic effects of individual molecules do not greatly affect the overall dynamics; this is estimated to be for populations above the range of 10^2 - 10^3 molecules per chemical species (36). For smaller numbers of molecules, the use of stochastic models may be more appropriate (204). Similarly, when modelling more global physiology, at a tissue or organ level for example, it is no longer feasible to describe the dynamics with such molecular detail and a more course-grained approach or multi-scale techniques are often necessary, e.g. (113, 130).

The biochemical mechanisms by which even relatively well-known enzymatic networks operate are not always clear (207). There are often multiple plausible network topologies that are consistent with the known underlying biology, and in this context, the idea of modularity within biology is important as it provides a natural means of iteratively building up a model description alongside successive biological experiments, with each half of the process informing the other. Model hypotheses can thus be constructed and compared with one another by incorporating new components into an existing model and then assessing the model based on the data (203). Recent advances in understanding the molecular origin of circadian rhythms in *Arabidopsis thaliana* (127, 128) offer an excellent example of this type of approach, whereby differential equation models have been able to offer predictions that have subsequently been tested experimentally (126).

Mechanistic modelling is a first step towards characterising the aetiology of many different types of diseases that are thought to be the result of disruptions to the normal functioning of signalling pathways and biochemical networks. Obtaining a detailed mechanistic understanding of such enzymatic control processes could have major implications in the study and potential treatment of disease at a molecular level. The circadian clock, for example, appears to play a central role in the physiology of plants and mammals, controlling many important cellular functions and influencing pathways implicated with a variety of diseases (192); indeed the timing of drug administration relative to circadian rhythms appears to strongly affect the efficacy of certain anticancer

agents (123). Signalling pathways are also strongly implicated in the origin of many cancers (177) and understanding the intricate networks of nonlinear interactions is key to being able to predict the impact of biochemical changes within the system, whether natural or artificially induced with the use of drugs (25).

Current investigations into biochemical networks are characterised by complex nonlinear dynamics, as well as by measurement and model uncertainty. When studying the possible structure of such systems, working hypotheses can be encoded as statistical models based on systems of nonlinear differential equations that capture all assumptions regarding the likely mechanisms of interaction, as we saw in the disease outbreak example in the previous chapter. As models increase in size and complexity, so too does the need for sophisticated statistical methodology that can consistently evaluate and update the evidence in favour of each model, as new data become available, see e.g. (196, 207). Given the limited and variable experimental data that is often available, a probabilistic approach based on Bayesian statistics offers a natural way of dealing with such parameter and model uncertainty. Rather than finding an optimal working set of parameters, as is the case in the frequentist setting (165), the Bayesian paradigm advocates averaging over all possible parameter sets with respect to their individual probabilities. This marginalisation procedure automatically provides a compromise between fitting the data and penalising model complexity, such that we may find the simplest model that is however still *complex enough* to accurately describe the observed dynamic behaviour.

Initial proof of concept investigations have shown that a Bayesian approach to model ranking can be very successful (203, 207), however it is acknowledged that performing Bayesian inference over ODE models is extremely challenging (27). The procedure is equivalent to evaluating integrals involving a highly nonlinear function over a high dimensional space. In more than 3 or 4 dimensions deterministic approaches are no longer feasible and we must resort to stochastic integration using simulation based Monte Carlo techniques. A common approach is to construct a Markov process that converges to the target posterior distribution (167), however this is not straightforward in practice due to strong correlation structures and expensive to compute likelihoods, as mentioned in the previous chapter.

In particular, the issue of identifiability is very important and has a direct impact on our ability to perform efficient frequentist or Bayesian statistical analysis. It has been

noted many times that ODE models of biochemical networks generally exhibit widely varying parameter sensitivities (22, 46, 61, 81, 197); investigation of second order sensitivities of these models, evaluated at the maximum likelihood, often reveals a wide eigenvalue spectrum which itself may change depending on the point in parameter space at which it is calculated. In settings with such varying parameter scalings, standard Markov chain Monte Carlo (MCMC) samplers generally have very poor mixing properties and produce highly correlated samples (77), resulting in estimates of the required Bayesian quantities with large Monte Carlo errors. This is often a result of structural unidentifiability of the model (165), such that parameters cannot be estimated with low variance. For example, if the output of a model depends strongly only on the ratio of two parameters θ_1/θ_2 , then the conditional distributions $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ may be well constrained, but their marginal distributions $p(\theta_1)$ and $p(\theta_2)$ might not be.

The differential geometric MCMC methods developed in this thesis seem particularly well suited to this type of application, since they exploit the natural representation of the model parameter space as a Riemannian manifold (32), which is induced using the Expected Fisher Information as a metric tensor (77). The Expected Fisher Information therefore *defines* a local distance measure and effectively allows the MCMC proposals to be based on the curvature of the manifold, which is directly defined by the parameter sensitivities of the underlying model describing the dynamical system, as we saw in Chapter 5.

We now examine the application of differential geometric MCMC methodology for performing Bayesian inference over ODE models of biologically relevant size with complex dynamics and partially unidentifiable structures. We perform posterior inference on two biologically realistic examples of biochemical networks that have recently been studied in the systems biology literature. We first consider a model describing the main circadian clock components in *Arabidopsis thaliana* (128), building up inference over different sets of model parameters to gain insight into the challenges of a statistical analysis. We perform inference using a synthetic dataset generated from the model and demonstrate how we may infer the posterior distributions over the parameters, in addition to obtaining low variance estimates of the marginal likelihoods for the purpose of model comparison. We then consider an alternative hypothesised model structure with a positive instead of negative feedback loop. We once again calculate marginal likelihoods based on the previous synthetic dataset, demonstrating the model ranking

procedure over these complex dynamical systems. Finally we consider a biochemical model based on mass-action kinetics with a transient response, which describes a cell signalling pathway of recent biological interest (17).

6.1.1 Circadian Rhythms in *Arabidopsis Thaliana*

Circadian rhythms play a central role in regulating the physiology of many living organisms. Indeed, many plants and animals have gained competitive advantages by adapting their internal physiology to synchronise with the environment they inhabit. Internal clocks have evolved to react to many external stimuli, with nutrients, light and temperature being obvious examples. The main mechanisms in a variety of different organisms that generate such rhythmic cellular activity appear to be commonly built upon positive and negative feedback loops. These 24 hour oscillations have to be sensitive enough to react to external changes appropriately, yet robust enough to remain stable in a variety of conditions over different timescales; interlocking feedback loops are a likely way of achieving these robustness qualities (150, 188).

Biological research communities have focused on a number of model organisms in order to study their basic design principles and increase the rate at which advances are made. *Arabidopsis thaliana* is widely accepted in plant biology as a model plant worthy of close and careful study (114). Although its genetic networks are much simpler than higher organisms, it is still complex enough to offer much insight into a large variety of biologically interesting mechanisms of biochemical interaction, such as negative feedback loops that induce nonlinear oscillatory behaviour. Overviews of the recent advances in understanding this reasonably complex biological system are given in (94, 132). Although this organism is not of immediate economic use, unlike other agricultural plants that have been studied, such as maize and rice (86), it is closely related to hundreds of thousands of other plants, which allows insights into its inner workings to be relatively easily extrapolated to these other species. There is a long history of study into this plant stretching back over the last one hundred years, although widespread adoption began in earnest in the 1980s, with the advent of gene-cloning methods and other advances in molecular biology and genetics (185). There are a number of practical considerations that have lead to the adoption of *Arabidopsis*. It is small and easily grown in laboratories with a short lifecycle of around two months. It is relatively easy to produce transgenic plants, with genetic modifications for the purpose of

probing various aspects of the plant's physiology, and in particular it is self-fertilising, capable of producing thousands of seeds from a single plant.

Circadian rhythms in *Arabidopsis* are generated from a central feedback loop producing cycles of mRNA and protein production and degradation (2). This small network comprises the transcription factors Circadian Clock Associated 1 (CCA1) and Late Elongated Hypocotyl (LHY), and the pseudo-response regulator Timing of Cab Expression 1 (TOC1). This network appears to operate within each individual cell, and such cell autonomy can be seen through experiments that induce oscillations with different phasing in different parts of the plant simultaneously (194). This distributed behaviour is also thought to help induce global robustness properties. As knowledge of this biological system has advanced, additional components and loops have emerged as also being important regulators of circadian rhythms. These include the transcription factors *Gigantea* (GI) (152), *Early Flowering 4* (ELF4) (56), *Zeitlupe* (ZTL) (107) and *Pseudo-response Regulators 3,5,7,9* (PRR3/5/7/9) (62, 138). Recent work has also investigated widespread circadian control throughout different parts of the plant, with regulation of circadian rhythms in the roots for example being controlled via sucrose levels (95), implying that while circadian rhythms can be organ specific, they do not appear to be organ autonomous. Overviews of the recent advances in understanding this complex biological system are given by McClung (132) and Hubbard et al. (94)

A number of mechanistic models based on ordinary differential equations have been proposed and analysed over the last 10 years, and such computer modelling has been an important tool for advancing knowledge of circadian genetic networks in *Arabidopsis*, however as yet there has been no use of more rigorous Bayesian statistical approaches to prediction and model comparison for describing the circadian genetic networks in *Arabidopsis*. Previous approaches have focused on parameter estimation using optimisation algorithms, and model discrimination has proceeded in a manually intensive way by comparing optimal predictions with experimental results under a variety of different environmental conditions.

Particularly interesting has been the computational modelling undertaken in the Millar Laboratory at the University of Edinburgh, who have developed increasingly complex mathematical models of the main feedback loop structures thought to be mainly responsible for driving circadian oscillations in *Arabidopsis* (126, 127, 128). Owing to a lack of data however, their approach to analysing these systems has consisted

of optimising the model parameters according to a cost function, constructed to take into account not only the data but also other qualitative features of the biological system, such as periodicity. Although useful insights have been gained in this way, the design of the cost function is somewhat ad hoc and as such there is no natural way of incorporating this scheme into a probabilistic framework to allow Bayesian model comparison.

As more data becomes available, hypothesis-driven Bayesian approaches will become more important and more widely used as a means of assimilating and updating current knowledge about a system in a consistent manner in light of new experimental data. The use of priors provides a natural means of incorporating information regarding appropriate reaction rate values, and the use of Bayes factors allows a means of comparing model structures, see e.g. (207).

The main challenge of performing Bayesian analysis over such large and complex statistical models based on systems of differential equations lies with the difficulty of sampling efficiently from the posterior distribution. We shall now implement the manifold MCMC methods developed in Chapter 3 to examine a large circadian model for *Arabidopsis*, and investigate the feasibility of performing a Bayesian analysis both at the parameter level and at the model level.

We employ a model based on the core circadian network in *Arabidopsis thaliana* comprising transcription factors LHY and CCA, and the pseudo-response regulator TOC1. As a simplification, both LHY and CCA are modelled as one component, since they have qualitatively similar behaviour. Michaelis-Menten kinetics are used to describe enzyme-driven protein degradation, with Hill functions describing transcriptional activation of mRNA for LHY/CCA and TOC1. The model is from (128) and is the minimal description of the network describing circadian behaviour of *Arabidopsis*, which we simulate in constant darkness. It consists of 6 nonlinear differential equations and a total of 24 parameters. The concentrations of the species are represented by $[TOC1]$ and $[LHY]$ with subscripts m , c and n denoting mRNA, protein within the cytoplasm and protein within the nucleus, respectively. The Hill coefficients a and b are set to 1 and 2 respectively, based on evidence from the literature (128). There are therefore 22 free parameters to be inferred from the experimental data, where (n_i, g_i) are transcription rates, (m_i, k_i) are degradation rates, (p_i) are translation rates, and (r_i) are rates defining transport between the nucleus and cytoplasm. The structure of

this biochemical network is represented in Figure 6.1 and its equations can be written as

$$\frac{d[LHY]_m}{dt} = \frac{n_1[TOC1]_n^a}{g_1^a + [TOC1]_n^a} - \frac{m_1[LHY]_m}{k_1 + [LHY]_m} \quad (6.1)$$

$$\frac{d[LHY]_c}{dt} = p_1[LHY]_m - r_1[LHY]_c + r_2[LHY]_n - \frac{m_2[LHY]_c}{k_2 + [LHY]_c} \quad (6.2)$$

$$\frac{d[LHY]_n}{dt} = r_1[LHY]_c - r_2[LHY]_n - \frac{m_3[LHY]_n}{k_3 + [LHY]_n} \quad (6.3)$$

$$\frac{d[TOC1]_m}{dt} = \frac{n_2 g_2^b}{g_2^b + [LHY]_n^b} - \frac{m_4[TOC1]_m}{k_4 + [TOC1]_m} \quad (6.4)$$

$$\frac{d[TOC1]_c}{dt} = p_2[TOC1]_m - r_3[TOC1]_c + r_4[TOC1]_n - \frac{m_5[TOC1]_c}{k_5 + [TOC1]_c} \quad (6.5)$$

$$\frac{d[TOC1]_n}{dt} = r_3[TOC1]_c - r_4[TOC1]_n - \frac{m_6[TOC1]_n}{k_6 + [TOC1]_n} \quad (6.6)$$

We used the following parameter values from the literature (128): $p_1 = 9.0002$, $p_2 = 3.6414$, $r_1 = 5.6429$, $r_2 = 8.2453$, $r_3 = 1.2789$, $r_4 = 5.3527$, $n_1 = 0.6187$, $n_2 = 7.7768$, $g_1 = 3.7051$, $g_2 = 9.7142$, $k_1 = 7.8618$, $m_1 = 7.3892$, $k_2 = 3.2829$, $m_2 = 0.4716$, $k_3 = 6.3907$, $m_3 = 4.1307$, $k_4 = 1.0631$, $m_4 = 5.7775$, $k_5 = 0.9271$, $m_5 = 4.4555$, $k_6 = 5.0376$, $m_6 = 7.6121$. The Hill coefficients were fixed at $a = 1$ and $b = 2$, and the initial conditions used for generating the synthetic data were $[LHY]_m(0) = 0.1290$, $[LHY]_c(0) = 13.6937$, $[LHY]_n(0) = 9.1584$, $[TOC1]_m(0) = 1.9919$, $[TOC1]_c(0) = 5.9266$, $[TOC1]_n(0) = 1.1007$.

6.1.2 Cell Signalling Networks

The modelling of cell surface receptors is an important means of furthering our understanding of intra-cellular signalling. Such mechanisms allow extra-cellular cues to drive activation and dynamic behaviour within each cell (179). In particular, changes in the environment are often encoded at a molecular level in terms of changes in the concentration of ligands outside the cell; such ligands may then bind to the surface of cells inducing internal changes brought about by phosphorylation mechanisms. A recent example of such a model is given in (17), which describes the nonlinear dynamic behaviour of the Erythropoietin (Epo) receptor in response to changes in the ligand

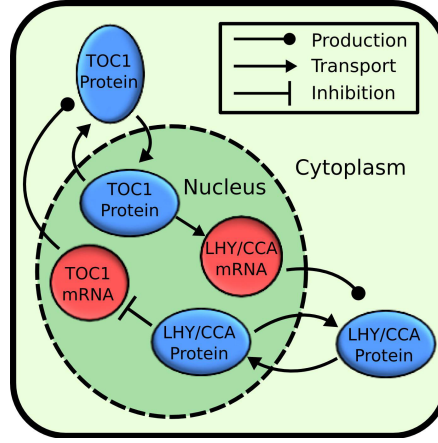


Figure 6.1: A representation of the model developed in (128), which we employ to model the main circadian oscillator network in *Arabidopsis thaliana* - TOC1 mRNA produces TOC1 protein in the cytoplasm, which is transported into the nucleus and increases production of LHY/CCA mRNA. This in turn, produces LHY/CCA protein in the cytoplasm, which is transported into the nucleus and inhibits production of TOC1 mRNA. This negative feedback loop is capable of producing oscillatory and highly nonlinear dynamical behaviour, providing a challenge for any optimisation or MCMC algorithm.

Epo concentrations outside the cell. Encoded signals sent in this manner have significant biological knock-on effects, invoking cellular responses that activate a number of signalling networks further downstream. The model we employ consists of 6 nonlinear differential equations, based on mass-action kinetics, with 8 parameters.

$$\begin{aligned}
 \frac{d[Epo]}{dt} &= -k_{on}[Epo][EpoR] + k_{on}k_D[Epo_EpoR] + k_{ex}[Epo_EpoR_i] \\
 \frac{d[EpoR]}{dt} &= -k_{on}[Epo][EpoR] + k_{on}k_D[Epo_EpoR] + k_t B_{max} \\
 &\quad -k_t[EpoR] + k_{ex}[Epo_EpoR_i] \\
 \frac{d[Epo_EpoR]}{dt} &= k_{on}[Epo][EpoR] - k_{on}k_D[Epo_EpoR] - k_e[Epo_EpoR] \\
 \frac{d[Epo_EpoR_i]}{dt} &= k_e[Epo_EpoR] - k_{ex}[Epo_EpoR_i] - k_{di}[Epo_EpoR_i] \\
 &\quad -k_{de}[Epo_EpoR_i] \\
 \frac{d[dEpo_i]}{dt} &= k_{di}[Epo_EpoR_i] \\
 \frac{d[dEpo_e]}{dt} &= k_{de}[Epo_EpoR_i]
 \end{aligned}$$

The parameter values (in \log_{10} scale) were determined from the literature to be:

$k_{on} = -4.091$, $k_{ex} = -2.447$, $k_t = -1.758$, $k_e = -1.177$, $k_{di} = -2.730$, $k_{de} = -1.884$. The remaining parameters were fixed to enforce identifiability: $k_D = 2.583$, $B_{max} = 2.821$. The initial conditions were fixed at: $[Epo](0) = 2098.9$, $[EpoR](0) = 662.22$, $[Epo_EpoR](0) = 0$, $[Epo_EpoR_i](0) = 0$, $[dEpo_i](0) = 0$, $[dEpo_e](0) = 0$.

We reparameterise the model such that we have parameters measured in \log_{10} space, allowing us to more easily consider a wide range of possible parameter values that vary by orders of magnitude. This model is complicated further by the addition of an observation model, since the biochemical species in this network are not all observable individually; in particular only the total concentration of $[Epo]$ and $[dEpo_e]$ is biologically observable, along with $[Epo_EpoR]$, such that our observations take the form $y_1 = [Epo] + [dEpo_e]$ and $y_2 = [Epo_EpoR]$.

6.1.3 Parameter Identifiability

A model parameter can be considered only weakly identifiable if changes in its value have very little effect on the output of the model. There are two main causes of unidentifiability (165); the first comes from measurement uncertainty in the data, and the second is a result of the model structure, in particular the mathematical relationship between the parameters and the states. Structural identifiability issues can occur, for example, when the parameters of a statistical model appear as a ratio; if the output of the model depends only on this ratio, then the numerator and denominator may effectively take on an infinite number of values. This effect has been termed parameter evaporation (197) since these unconstrained parameters can tend to infinity. This poses a particular problem for methods that make use of 2nd order geometric information, as the unconstrained parameters can result in an ill-conditioned Hessian or Fisher Information matrix (200), and so the point-wise variance estimates of such parameters can be very poor.

In the context of the circadian biochemical equations above, the commonly used Michaelis-Menten terms are of particular interest (197). This is a model to describe an irreversible enzymatic reaction and relates the rate of the reaction to the concentration of a substrate, $[S]$. Let us consider the Michaelis-Menten equation,

$$r = \frac{r_{\max}[S]}{K + [S]} \quad (6.7)$$

where r is the current reaction rate, r_{\max} is the maximum reaction rate, and K is the Michaelis-Menten rate, which is inversely related to the enzyme's affinity for the substrate S . Firstly let us note that,

$$K \ll [S], \quad r \approx r_{\max}, \quad \text{since } \frac{[S]}{K + [S]} \approx 1 \quad (6.8)$$

In this case, when the rate constant K is much less than the concentration level of S , the overall reaction rate is determined by r_{\max} . K is therefore unidentifiable as it has almost no effect on the model output. Similarly we see that,

$$K \gg [S], \quad r \approx \frac{r_{\max}[S]}{K} \quad (6.9)$$

and so it is the ratio of r_{\max} to K that is important in determining model behaviour. In this case, both r_{\max} and K are potentially unidentifiable, depending on the values of $[S]$. This unidentifiability results from our decision to employ Michaelis-Menten kinetics to describe our system; the fact that parameter values could blow up to infinity or evaporate to zero may not have any bearing on the biological reality, but are more likely to be artefacts of our model approximation of the underlying system. Within a Bayesian setting, the use of priors may be used to enforce weak identifiability and improve numerical conditioning (200).

6.2 Implementation

We examine the application of differential geometric MCMC methodology for performing Bayesian inference over large ODE models that exhibit complex dynamics and partially unidentifiable structures. We embed the MCMC samplers within a population framework to help escape local maxima and fully explore the parameter space, noting that the samples obtained can then be further used for calculating Bayes factors via thermodynamic integration (27, 67, 118). Within this framework, we explore multiple tempered distributions simultaneously, each defined by a power posterior (67). These form a smooth family of distributions between the prior and posterior, and exchange moves between these distributions allow faster convergence to the global mode of interest. We employ the same temperature schedule as detailed in (29) and more details are given in Section 3.5.

For each of the subsets of parameters, a burn-in phase of between 10,000 and 20,000 samples was found to be sufficiently long for the Markov chains to converge to the stationary distributions defined by each of the power posteriors used. For inferring the complete set of parameters, the burn-in phase lasted 20,000 iterations and 100,000 samples were collected from each temperature for the purpose of estimating the marginal likelihood. During this time the step sizes of the parameters were adjusted every 100 iterations to achieve an acceptance ratio of between 20% and 50% for standard Metropolis and between 30% and 70% for manifold sampling using simplified mMALA. After the burn-in period, step sizes were fixed to ensure samples were drawn from the stationary distribution.

The cell signalling model is made identifiable by fixing 2 parameters, the values of which may be obtained experimentally (17). The initial condition of $EpoR$ may also be found by experiment (17). The initial condition of Epo is assumed to be known, and the other four initial conditions are set to zero. The initial conditions of the observed species $Epo + dEpo_e$ and Epo_EpoR are therefore fully known. The Fisher Information for the remaining 6 parameters is well-conditioned and we may therefore infer these together, without using a blocking approach. For this model, a burn-in phase of 5000 samples was found to be sufficiently long for the Markov chains to converge.

We obtained the auxiliary sensitivity equations for our ODE models using the Symbolic Math Toolbox in Matlab. This automated process is extremely fast and helps prevent human mathematical errors. We solve them making use of the SBToolbox2 (178) for Matlab, which provides an interface to a C implementation of the Sundials solver (90) and is up to 2 orders of magnitude faster than using the built-in solvers in Matlab.

Synthetic datasets are very useful for evaluating the performance of new methodology, and with complex ODE models it is convenient to have a set of parameter values that generate the data as a benchmark. We investigate the models using a variety of noise levels, which we detail along with the results in the next section; ultimately we wish to test our methodology using data with biologically realistic numbers of observations and levels of noise. For the circadian model we generate data by sampling 16 data points for each species evenly over a time period of 48 hours, using parameter values given in the literature (128). Using a Gaussian error model, we add zero mean noise to the data points generated for species $[LHY]_{[m,c,n]}$ and $[TOC1]_{[m,c,n]}$, with

standard deviation proportional to the amplitude of the concentration of each species. The circadian model we investigate has the difficulty that the Fisher Information over all parameters is numerically ill-conditioned, as often happens when parameters are weakly identifiable (200).

6.2.1 The Choice of Priors and Ill-Conditioned Metric Tensors

Priors are used in Bayesian statistics as a means of incorporating existing knowledge regarding likely parameter values. In this setting, such knowledge commonly arises from consideration of the underlying biology and experimental observations, however it may also arise by considering the model itself. In previous work examining oscillatory behaviour of biochemical networks (89), bifurcation analysis of a relatively simple delay differential equation model was employed to find the regions of parameter space that permitted oscillatory output, and this information then informed the choice of priors. This approach however is obviously dependent on the model being amenable to such analysis; bifurcation analysis on much larger, more complex models quickly becomes unfeasible. The idea of informing the range of admissible parameter values based on a mathematical analysis of the model has also been suggested in (197), in particular for the case of examining Michaelis-Menten kinetics, where the aim was to avoid numerical issues and parameter evaporation. Using such an approach for the circadian rhythm models considered in this chapter, we can enforce weak identifiability without restricting the range of behaviour our model can produce. In this case we want K to be neither too small nor too big, but rather roughly the same order of magnitude as we would expect $[S]$ to be.

In the clock model we employ vague gamma priors for the parameters, with shape parameter $k = 1.5$ and scale parameter $\theta = 10$. This prior results in the canonical parameters having positive support, reflecting the fact that parameters correspond to physical rate constants, and it discourages parameter values approaching zero, reflecting the assumption that all terms in the model should play a role in the overall biochemical process. Such priors are also suitable for preventing the rate parameters becoming too large, since we assume that all chemical processes are slow enough to be observed. The prior thus serves to constrain the parameter values, which has the added numerical benefit of regularising the Fisher Information, when it might otherwise be near singular with a very high condition number (200), as we note that it is often beneficial to add

a diagonal scaling to the metric tensor to improve numerical conditioning. The overall metric tensor is formed by taking the Expected Fisher Information and subtracting the negative Hessian of the prior, as we saw previously in Chapter 4.

In the signalling model we employ Gaussian priors with a mean of -2 and a standard deviation of 2; since we reparameterise the model in \log_{10} space we no longer require our prior to have only positive support, and such priors cover parameter values of many orders of magnitude.

6.3 Circadian Model Results

We first consider three subsets of the parameters individually; the linear parameters $(\mathbf{p}_{1:2}, \mathbf{r}_{1:4})$, the transcription parameters $(\mathbf{n}_{1:2}, \mathbf{g}_{1:2})$, and the Michaelis-Menten parameters $(\mathbf{m}_{1:6}, \mathbf{k}_{1:6})$. This allows us to see the potential challenges of inferring parameters in these three different types of functions. We perform inference over each set of parameters with the other sets of parameters fixed at their true values.

6.3.1 Linear Parameters

For now, we run a single chain initialised on the true parameter values for the purpose of evaluating the sampling properties at the mode, and we fix the initial conditions.

We first compare posterior sampling of the linear parameters using Metropolis, Simplified mMALA, MALA and RMHMC, and a comparison of the sample traces is given in Figure 6.2. We employ a component-wise Metropolis sampler, whereby each parameter is updated sequentially conditioned on all other parameters using a Gaussian proposal, and we note that the component-wise Metropolis-Hastings sampler therefore requires the ODE model to be solved once for each parameter. The proposal variances are automatically tuned to achieve acceptance rates of between 20% and 50%, and then fixed before drawing posterior samples. This standard sampler performs very poorly and is unable to sample efficiently from the strongly correlated posterior distribution. The manifold methods however perform far better, making use of the local geometry information to propose better moves through parameter space. RMHMC makes proposals that follow geodesic paths across the manifold, however in our implementation it

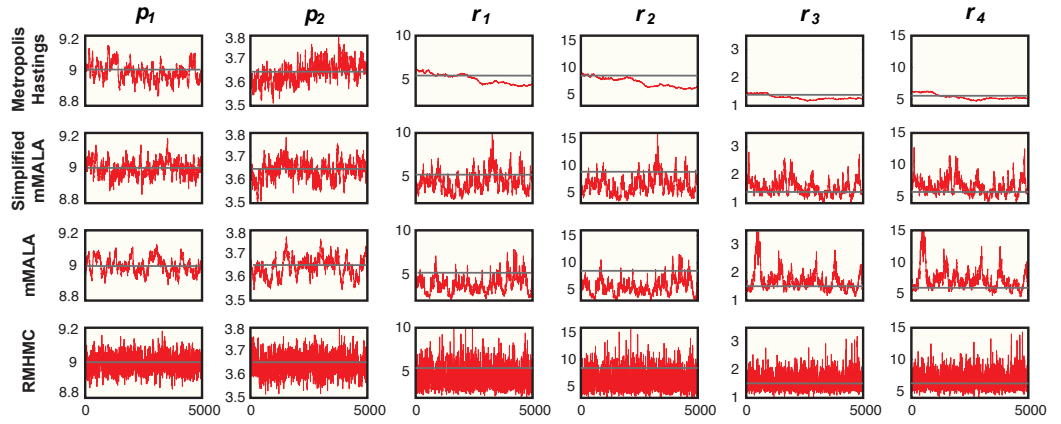


Figure 6.2: Comparison of posterior samples of the linear parameters in the circadian model obtained using a variety of MCMC sampling algorithms - The true parameter values are represented by the grey line. The woeful performance of the Metropolis-Hastings sampler is apparent from the slowly moving, highly correlated samples it produces. In contrast, the samplers that make proposals using local geometric information, as defined by the sensitivities of the ODE system, do far better. RMHMC produces nearly uncorrelated samples, however this is at great computational expense since the 2nd order sensitivities must be calculated multiple times over the geodesic proposal path. Similarly mMALA requires the 2nd order sensitivities to be computed. Simplified mMALA offers performance approaching that of mMALA, at a much reduced computational cost requiring only the 1st order sensitivities to be calculated.

is computationally much more expensive¹, requiring an average of five evaluations of the 2nd order sensitivity equations and an average of twenty-five evaluations of the 1st order sensitivity equations. mMALA makes moves based on a diffusion process across the manifold, and also performs much better than Metropolis-Hastings, although again this is computationally expensive because of the need to calculate second order sensitivities of the differential equation model; it requires two evaluations of the 2nd order sensitivity equations per iteration. Simplified mMALA makes valid MCMC proposals based on an approximate diffusion across the manifold. It gives similar results to mMALA and is computationally much more efficient as it only requires two evaluations of the first order sensitivities per iteration. In practice, we find that solving the extended 1st order ODE system twice for simplified mMALA is only 3 to 4 times slower than solving the original system for each parameter in a component-wise Metropolis-Hastings sampler, yet it produces samples that are visibly far better (see Figure 6.2).

This verifies the results in Chapter 5, where we observed that simplified mMALA generally provided the best balance, using the local geometry to improve sampling, while remaining computationally feasible. We note that for our differential equation models the extended 1st order system consists only of additional linear equations that may be solved very efficiently (90). We therefore employ simplified mMALA as our sampler of choice for all subsequent experiments in this paper.

Figure 6.3 shows scatter plots of the samples obtained from simplified mMALA for each parameter combination. Even with these linear parameters, there are some very strong correlations and severe scaling differences, illustrating the difficulty of sampling from statistical models of this type. As an example, let us consider parameters r_1 and r_2 , which define transport between the nucleus and cytoplasm for LHY/CCA. There is a very strong correlation structure between these parameters; as the concentration of the protein in the nucleus increases, the concentration of protein in the cytoplasm decreases, and vice versa. There is a similar strong correlation structure between parameters r_3 and r_4 describing nuclear/cytoplasmic transport for TOC1. Although the other parameters do not exhibit such strong dependencies, their differing scales are evident, which further compounds the difficulties.

¹This performance could be substantially improved by using adjoint differentiation methods and programming using a compiled language, rather than the interpreted language Matlab, as we use here.

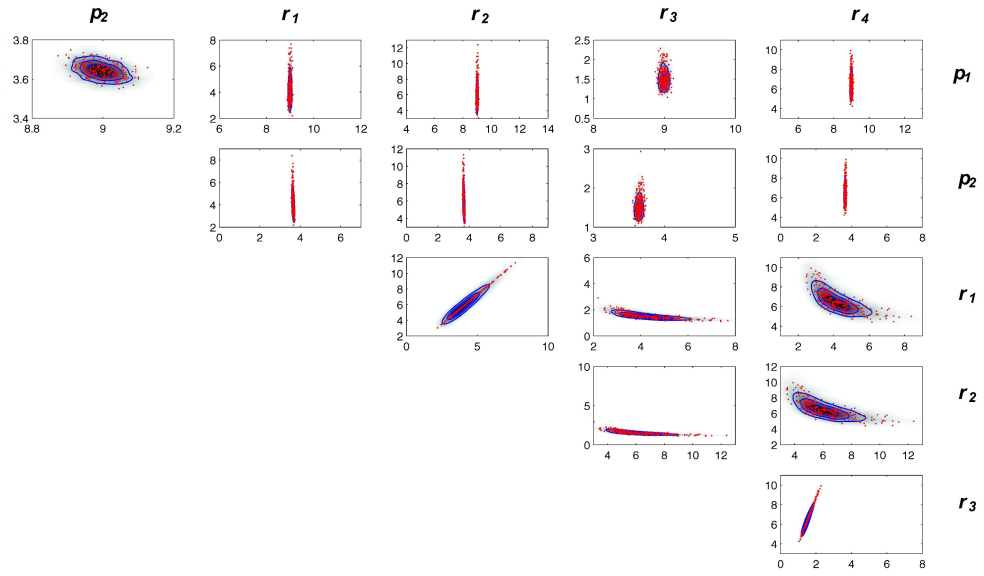


Figure 6.3: Pairwise scatter plots and density estimates of posterior samples of the linear parameters in the circadian model - We observe a strong correlation structure present in this nonlinear system for the linear parameters, with all other parameters fixed at their true values. The samples were obtained using the simplified mMALA sampler with a $\text{Gamma}(1.5, 2)$ prior, using artificially generated data with zero mean Gaussian noise, whose standard deviation was equal to 1% of the amplitude of the oscillations in each species.

We also compared the effect that changing the noise level has on parameter inference, and in particular on the practical identifiability of the model. Even with very low levels of noise, we observed wide variability in the parameter estimates, however the model was still able to make predictions with high certainty, as has been observed in biological models examined in previous work (197); dealing with large uncertainty in parameter values is an unavoidable part of the modelling process. For many models the experimental accuracy required to obtain tight bounds on individual parameters is simply unachievable, and indeed often undesirable if we are mainly interested in the model predictions. The careful use of priors can therefore help to constrain parameter values to biologically meaningful ranges.

Finally we examine the effect of the prior on the posterior distribution, using a gamma prior over the parameters with selection of scale parameters, $\theta = [2, 10, 20]$. We observe that the choice of prior in this case has little effect on identifiable parameters, however there is a noticeable effect on those parameters that are unidentifiable; such parameters may take on a wide range of values while the model output still describes the data, see Figure 6.4.

6.3.2 Transcription Parameters

We now consider inference over the 4 transcription parameters with Hill coefficients, and fix all other parameters at their true values. Again, there are strong correlations between these parameters, particularly between parameters appearing together within the same algebraic term. This correlation structure is shown in Figure 6.5 and was seen to remain even with increasing levels of error variance in the data.

6.3.3 Michaelis-Menten Parameters

Fixing all parameters apart from those appearing in the Michaelis-Menten terms, we investigate the effect of bounding the eigenvalues of the metric tensor in order to improve numerical conditioning and obtain better sampling. Based on the analysis of the Michaelis-Menten terms in Section 6.1.3, we may note that each pair of parameters appears together in a Michaelis-Menten term and in certain cases the system may respond to changes in the ratios of these pairs. This can result in the Fisher Information becoming very badly conditioned (condition number $> 10^8$) and the resulting numerical inaccuracy results in very small step sizes and highly correlated samples being drawn.

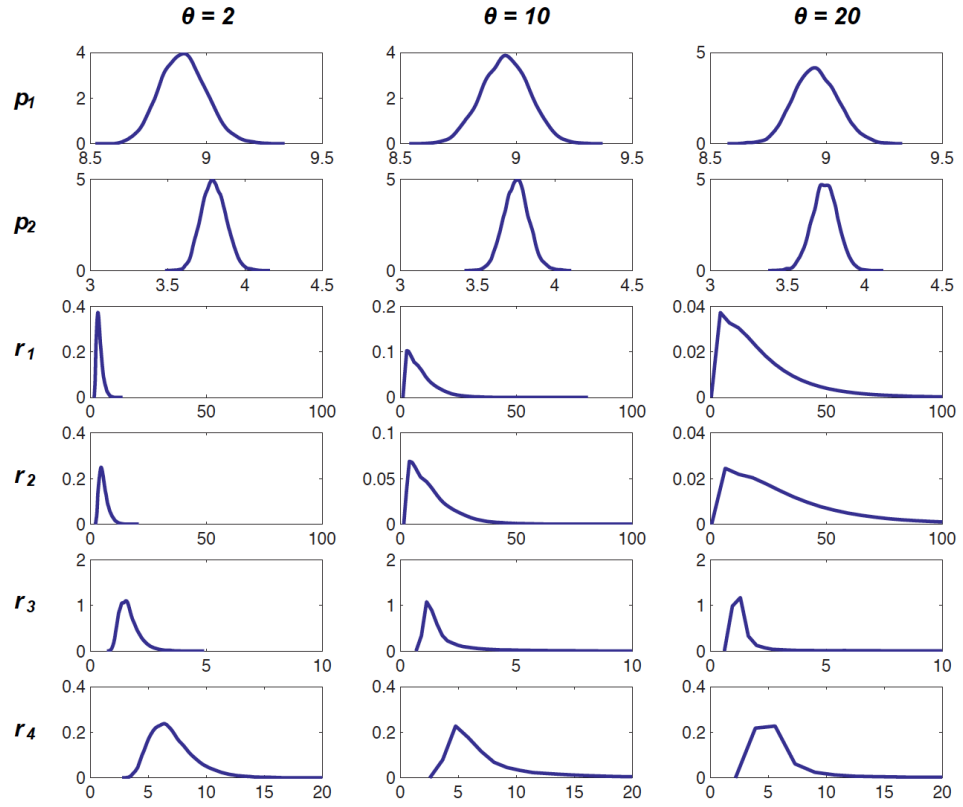


Figure 6.4: Posterior distributions over the linear parameters of the circadian model using different priors - The samples were obtained using simplified mMALA with $\text{Gamma}(1.5, 2)$, $\text{Gamma}(1.5, 10)$ and $\text{Gamma}(1.5, 20)$ priors. We observe that the prior only appears to have an effect on the unidentifiable parameters.

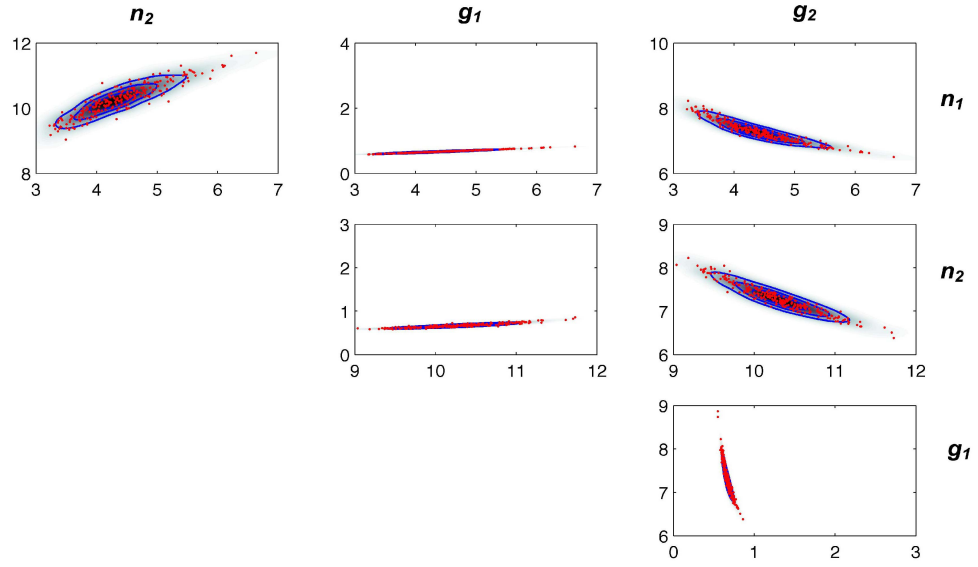


Figure 6.5: Pairwise scatter plots and density estimates of posterior samples of the Hill parameters in the circadian model - The samples were obtained using simplified mMALA with a $\text{Gamma}(1.5, 2)$ prior. This time the Gaussian noise used to generate the data had standard deviation equal to 10% of the amplitude of the oscillations. Even with this higher level of uncertainty in the data, the samples still exhibit very strong correlation structure, reinforcing the need for more advanced sampling methodology.

The condition number is given by the ratio of largest to smallest eigenvalues, and so by bounding the lowest eigenvalue of the metric tensor we improve the numerical conditioning, and this has a significant impact on sampling, as is shown in Figure 6.6. We may set a bound on the lowest eigenvalue by performing a singular value decomposition (SVD) of the metric tensor, and setting a lower bound on the eigenvalues appearing in the diagonal matrix. We can consider the SVD as a method of obtaining an alternative basis for the local vector (tangent) space at a given point, where the eigenvalues give the magnitude of the bases in each direction. We see that inverting the decomposed metric tensor simply involves inverting the diagonal matrix of the eigenvalues, and so the lower bound becomes an upper bound on the variance of the proposals, since it is the inverse of the metric tensor that is employed as the covariance of the proposal step in mMALA. This therefore helps to prevent the numerical instability that is seen to occur when there are unidentifiable parameters. Figure 6.6 shows the traces of Michaelis-Menten parameter samples obtained with and without bounding the lowest eigenvalue, and this approach can be seen to improve the sampling of weakly identifiable parameters.

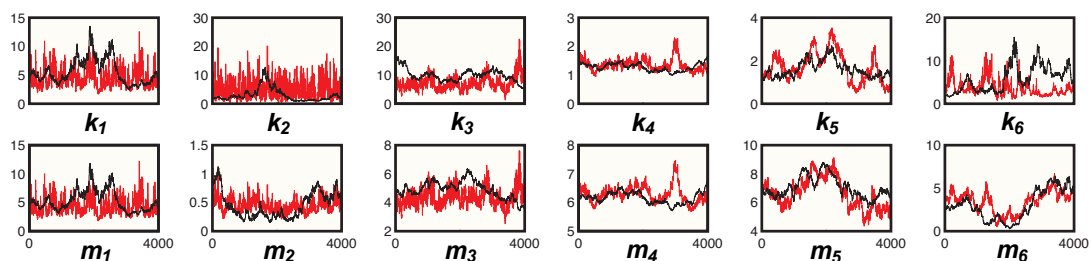


Figure 6.6: Comparison of sampling Michaelis-Menten parameters, with and without a bounded eigenvalue strategy - The 12 Michaelis-Menten parameters from the circadian model were sampled using simplified mMALA, with and without a bounded eigenvalue strategy, displayed on the red and black respectively. All 12 parameters were sampled simultaneously. In particular, Parameters (m_1, k_1) are only weakly identifiable and this results in an ill-conditioned Fisher Information that negatively impacts on the sampling efficiency using the full Fisher Information matrix. By performing a singular value decomposition and bounding the lowest eigenvalue, better conditioned FI matrices can be obtained; this numerical stability subsequently allows for more efficient sampling using a bounded eigenvalue approach.

Scatter plots of the parameter pairs (m_i, k_i) are also shown in Figure 6.7. Interest-

ingly, we see that the pair of Michaelis-Menten parameters with the strongest correlation (m_1, k_1) is indeed for a species with low concentration, such that $k_1 \gg [LHY]_m$. We see that this type of severe correlation structure was premeditated in Section 6.1.3, where we considered the case of the *ratio* of m_1 to k_1 driving the dynamics. Clearly, when performing inference using MCMC, we wish to make moves in parameter space that take into account these strong correlations, and the differential geometric MCMC approaches developed in Chapter 3 do this automatically.

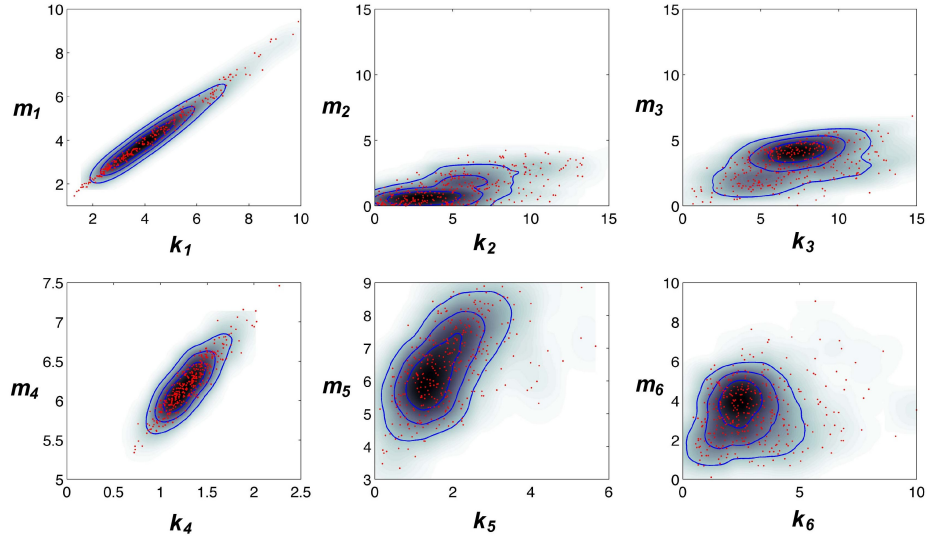


Figure 6.7: Scatter plots and density estimates of the pairs of Michaelis-Menten parameters from the circadian model - The samples were obtained using simplified mMALA with a $\text{Gamma}(1.5, 2)$ prior and Gaussian noise with standard deviation equal to 10% of the amplitudes of oscillation. The possibility of such strong correlation structure can be predicted by an analytic consideration of the underlying equations.

6.3.4 Inference with Full and Partial Observations

Until now we have employed a single chain initialised at the correct parameter values to examine the local mixing properties of manifold MCMC methodology. We now consider the effect of using random starting values for our parameters. Performing inference on just the linear parameters, keeping all others fixed, we find that there appears to be a single mode, which our Markov chain can reach regardless of the starting parameters. However, if we infer both the linear parameters and the transcription parameters, we

find there is another mode that our Markov chain may converge to. This alternative set of parameters gives a non-oscillatory output that roughly corresponds to the average of the true model output, as shown in Figure 6.8. Systems with complex nonlinear dynamics are particularly susceptible to this problem, since local maxima may occur when the model moves in and out of phase with different parts of the data. Any MCMC methodology used for this problem must therefore not only have good *local* mixing properties, but it must also be capable of making more *global* steps, allowing it to escape from local modes of negligible probability mass. Using simplified mMALA to explore a population of tempered distributions (27) allows us to resolve this issue.

We infer all model parameters and initial conditions together based on all states being observed, then with only the 4 protein states observed, and finally with only the 2 mRNA states observed. Each time we employed a population scheme with 50 tempered distributions and the eigenvalues of the metric tensor were bounded to improve numerical conditioning. Figure 6.9 shows the differences in the predictive model outputs for differing numbers of observed states. Such simulations are important as it is usually not possible to obtain measurements for all components in a biological system, due to either financial or technical constraints.

6.3.5 Estimating Marginal Likelihoods for Model Ranking

Finally, we demonstrate that marginal likelihoods may be estimated with low variance using this combination of population MCMC and differential geometric sampling, via the use of thermodynamic integration (67, 118). We show how this approach may be used for ranking multiple model hypotheses encoded as systems of differential equations, even for more complex systems with larger numbers of parameters. We infer all parameters and initial values over 50 tempered distributions employing the lower bounded eigenvalue strategy to improve numerical conditioning of each metric tensor and upper bound the proposal variance. We drew 20,000 samples as burn in, then stored the next 100,000 samples for each temperature. Marginal likelihood estimates were calculated from the resulting samples based on a trapezoidal approximation to the thermodynamic integral in a same manner as (27). The simulations were repeated 10 times and the summary statistics are presented in Table 6.1. Stable, low variance estimates of the marginal likelihoods were obtained for the fully observed and the two partially observed systems.

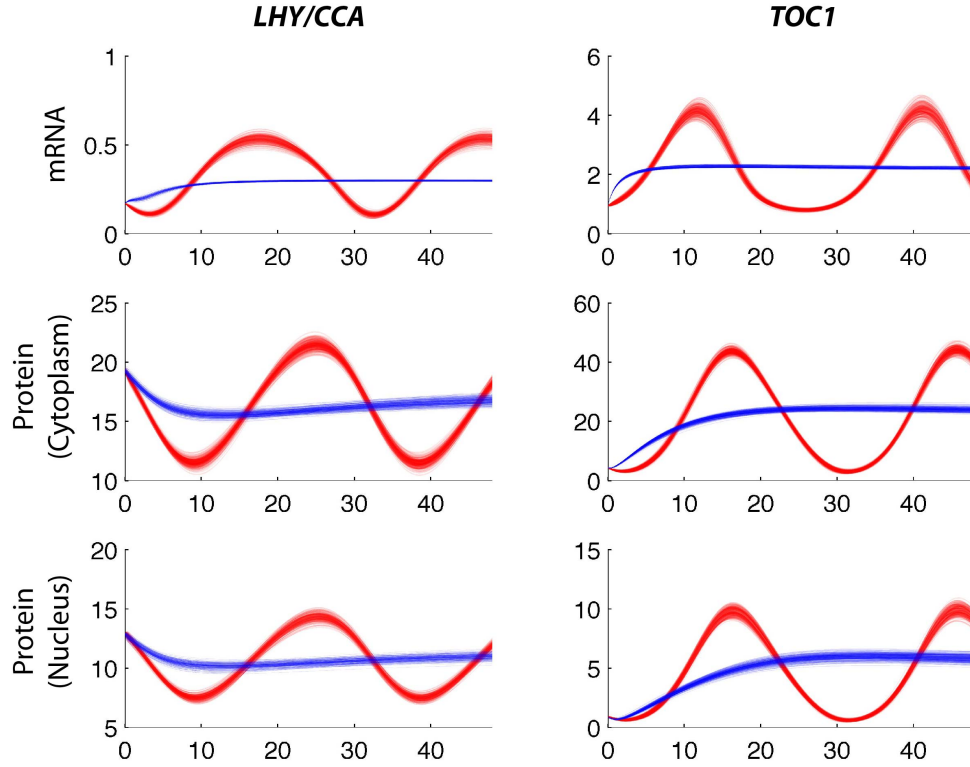


Figure 6.8: Population-based MCMC is extremely useful for ensuring convergence to the correct mode (27) - As an illustration, we first employ a single chain initialised at a chosen set of parameter values to obtain samples from the correct posterior mode using a manifold MCMC method. The predictive model output based on these samples is shown in red. We then initialised a single chain at a random set of parameter values sampled from a $\text{Gamma}(1.5, 2)$ prior distribution, and ran it until apparent convergence. This time the predictive model output converges to a local maximum, shown in blue, with the predictions cutting halfway through the oscillations in the data. Such local modes may also occur as the model output moves in and out of phase with oscillatory data.

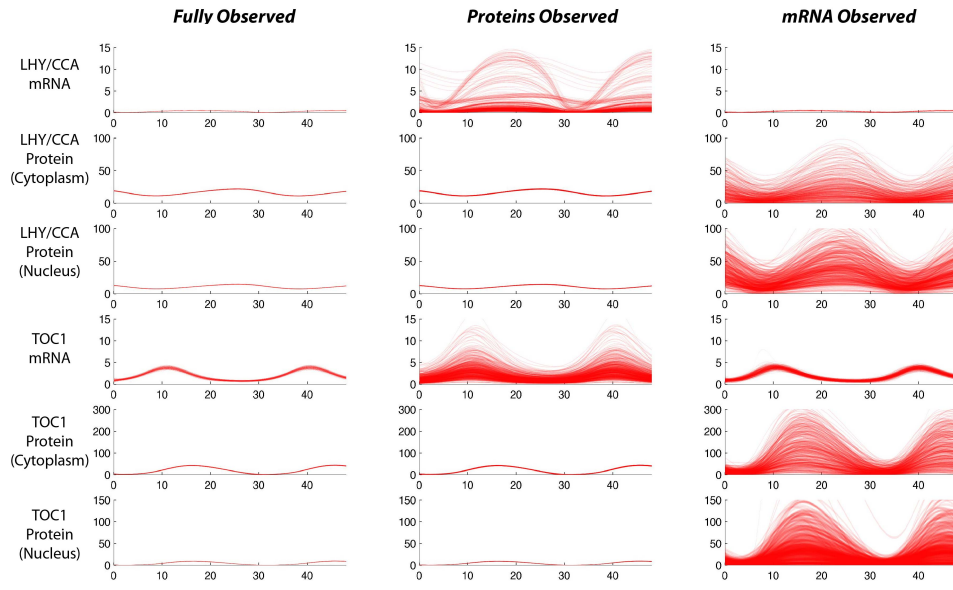


Figure 6.9: Comparison of posterior outputs from the circadian model with unobserved species - Samples were obtained by performing Bayesian inference over all parameters using a population sampling scheme with simplified mMALA. For a fully observed system, the posterior model predictions are all made with low uncertainty. In contrast, we obtain much more vague predictions for unobserved species, and the uncertainty in the predictions increases as the number of observed species decreases.

Using the same data we consider an alternative model hypothesis which we create by replacing the negative feedback loop in this circadian model (see Figure 6.1) with a positive feedback loop. This requires only a very small change to the underlying equations; in the equation describing the change in concentration of $[TOC1]_m$ we simply replace the term $\frac{n_2 g_2^b}{g_2^b + [LHY]_n^b}$ with $\frac{n_2 [LHY]_n^b}{g_2^b + [LHY]_n^b}$. The summary statistics for this new model are given in Table 6.2. We see that such a simple change must drastically alter the possible range of dynamic behaviour of this circadian model, since the marginal likelihoods are now far smaller. The Bayes factors suggest very strong support in favour of the model with a negative feedback loop, and we conclude that the positive feedback loop model is unable to reproduce the observed oscillatory behaviour. Indeed if we compare the posterior model outputs of the two models, shown in Figure 6.10, we see that the positive feedback model is unable to adequately describe the observed data.

Table 6.1: Marginal likelihood estimates for each of the sets of synthetic observations for the circadian ODE model with negative feedback loop

Observed States	Mean	Standard Deviation
Full	-119.3	0.7
Protein	-142.4	0.9
mRNA	6.7	0.4

Table 6.2: Marginal likelihood estimates for each of the sets of synthetic observations for the alternative circadian ODE model based on a positive feedback loop

Observed States	Mean	Standard Deviation
Full	-1567.3	13
Protein	-1443.7	12.7
mRNA	-61.7	0.5

6.4 Cell Signalling Model Results

We now infer parameters over the cell signalling model in which the observations are a linear combination of the underlying modelled species, demonstrating that manifold sampling methodology extends straightforwardly to incorporate observation models. It also demonstrates how this methodology may be used to infer parameters of models

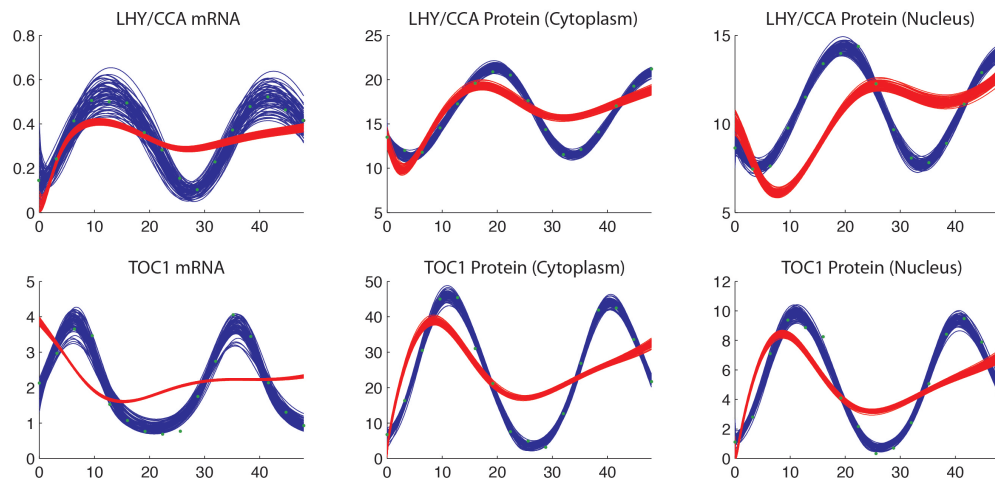


Figure 6.10: Comparison of posterior outputs from two fully observed circadian models - Samples were obtained from 2 circadian models by performing Bayesian inference over all parameters using a population sampling scheme with simplified mMALA. The data was generated from a circadian model with a negative feedback loop, resulting in oscillatory dynamical behaviour. The inferred posterior model predictions from the negative feedback model are shown in blue; the inferred posterior model predictions from the positive feedback model are shown in red. It is clear that the positive feedback model is unable to produce the observed oscillatory behaviour, as was suggested by the relative marginal likelihoods of the two models.

based on commonly used mass-action kinetics. The model is parameterised in \log_{10} space allowing exploration over several orders of magnitude. Despite having only 6 parameters, this cell signalling model still exhibits reasonably complex dynamics and sensitivities that change markedly throughout the parameter space. This is shown in Figure 6.13, where we observe the path taken by a Markov chain during the burn-in period guided by the sensitivities of the model parameters, which dramatically change as the chain explores new regions of the space. Despite the very wide priors covering orders of magnitude, the population-based simplified mMALA sampler is still able to converge to the correct mode. In addition, we find that such a model based on mass-action kinetics also induces a highly correlated, asymmetric posterior distribution that is challenging to sample from, as shown in Figure 6.12, particularly as the parameter values are of differing orders of magnitude. Finally, model predictions are made with very low uncertainty, Figure 6.11, however it has been noted (17) that a careful initial analysis of the model is necessary to achieve this structural and practical identifiability.

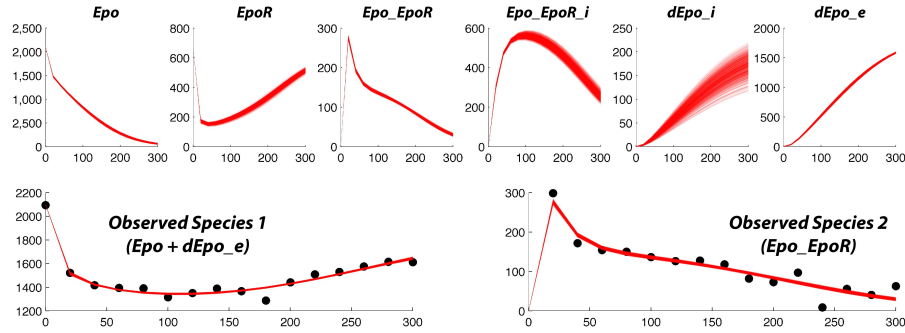


Figure 6.11: Posterior model predictions for the cell signalling model with additional observation model - The top row shows the posterior model predictions for each of the biochemical species in the Erythropoietin cell signalling model, and the bottom row shows the predictive model outputs using the additional observation model, with the dataset shown in black. The tightest predictions are made for *Epo*, *dEpo_e* and *Epo_EpoR*, and indeed the observations are a linear combination of these species. Inference on this model produces reasonably tight predictions for the unobserved species too, although this is a result of constraints on the model to ensure identifiability (17).

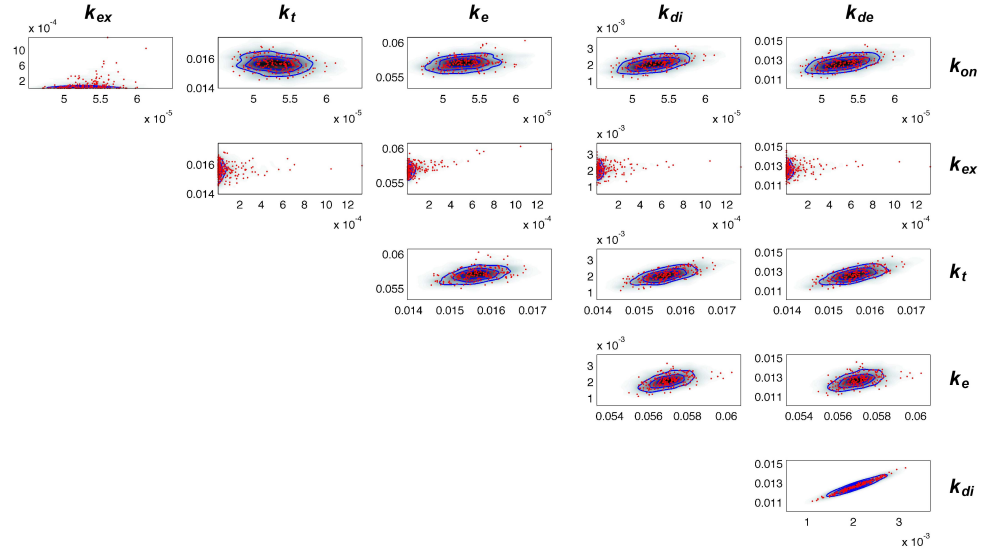


Figure 6.12: Scatter plots and density estimates of the posterior distribution of a cell signalling model - Scatter plots and density estimates of the parameter samples from the Erythropoietin cell signalling model obtained using simplified mMALA with a $\mathcal{N}(-2, 2)$ prior and Gaussian noise with standard deviation equal to 10% of the amplitudes of the outputs. We note the orders of magnitude difference in the scaling of each of the parameters.

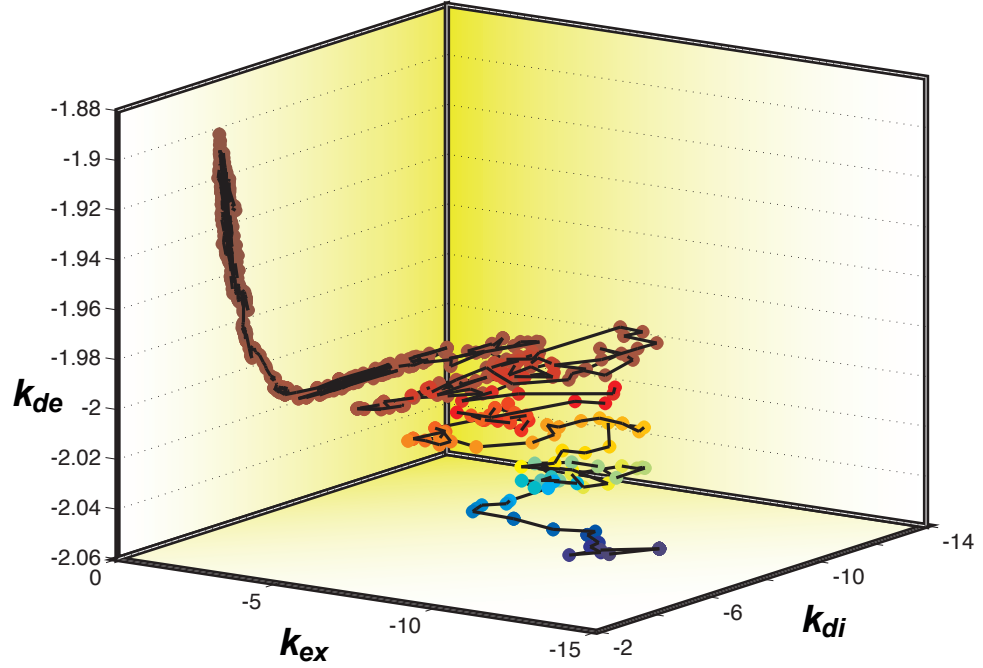


Figure 6.13: Example path taken during the burn-in phase of a Markov chain exploring the posterior distribution of the cell signalling model - The path shown was taken during the burn-in phase of a Markov chain as it moved through 3 of the 6 dimensions of parameter space of the cell signalling model, with the colour representing the posterior probability of the current set of parameters; the chain starts with very low probability (blue) moves to higher probability region (red) and finally finds the posterior mode (brown). All proposal steps were made using local sensitivity information via simplified mMALA sampling. Interestingly, the sensitivities of the parameters change dramatically as the chain moves through the parameter space. At the beginning of the path (shown by the blue points), the chain moves in all 3 of these parameter directions, until it reaches a plateau along which 2 of the parameters become tightly constrained. At the end of the plateau, the sensitivities of 2 of the parameters switch with the chain now moving upwards, with little movement along the previously preferred axis.

6.5 Conclusions

Statistical analysis of mechanistic models describing biological systems plays a vital role in furthering our understanding of the underlying mechanisms driving observed behaviour (17, 191, 196, 207). Both frequentist and Bayesian approaches to statistical inference face the challenge of searching through high dimensional parameter space; the former relying on optimisation, the latter on MCMC sampling.

In this chapter we have focused on the Bayesian approach and find standard sampling methods are inadequate for this task and are unable to sample even a subset of the total parameters efficiently. We require a sampler that takes into account the locally changing correlation structure in order to propose moves that are accepted with high probability, such that samples are drawn efficiently from the parameter space. One approach is to employ adaptive MCMC methods (82) that estimate the covariance using previous samples from the Markov chain, however the rate of diminishing adaptation must be carefully controlled to ensure convergence and it may also take a while to obtain reasonable estimates of what is essentially an estimate of the covariance structure; as adaptation diminishes, this estimate of the correlation structure becomes constant, which may not be appropriate for exploring spaces in which the local structure varies from point to point. In contrast, we obtain the local covariance structure *directly* at each point using the induced Riemannian geometry of the model via the Expected Fisher Information. Manifold MCMC methods (77) make moves in parameter space based directly on the sensitivities of the underlying differential equations with respect to each parameter, resulting in efficient convergence to the correct posterior distribution.

A further inferential challenge stems from the fact that many complex ODE models may induce multiple local modes, which can occur for example when a non-oscillatory steady state solution is found that may be fitted to the average values of the observed data as we have illustrated in Figure 6.8. Population MCMC approaches offer an effective solution to this problem and in combination with differential geometric sampling methodology allow for accurate estimation of marginal likelihoods (67), which can be used for model comparison.

The RMHMC sampler clearly offers the most effective scheme in terms of generating nearly-independent posterior samples, however solving the equations to calculate

the geodesic paths on the induced manifold can be computationally expensive. Ideas for mitigating this issue include the adoption of adjoint differentiation methods for efficiently calculating sensitivities of large numbers of parameters, the use of appropriate guiding Hamiltonians in the proposal mechanism (57), and implementation in computationally more efficient programming languages. Further work will bring these costs down and make RMHMC available to a larger class of problems. In the meantime, the simplified mMALA sampler offers the benefits of a geometric approach to MCMC, while remaining computationally feasible for larger models.

Many realistic biological models have parameters that are unidentifiable and this issue has recently received a lot of attention (165). Such problems can be addressed by considering a reduced model or, in a Bayesian setting, employing biological and mathematical analysis to inform the choice of priors and enforce weak identifiability (197). Although ODEs may be employed to model a wide range of important natural phenomena, it is worth noting that other modelling formalisms, such as stochastic and partial differential equations (141, 204), are also important in other settings. Recent work on calculating the Fisher Information for SDEs (112) suggest that such systems may also be considered within a differential geometric framework.

The manifold samplers we developed and employed use the local sensitivities of the model to construct an efficient Markov transition kernel, allowing us to sample from complex posterior densities, which in turn provide the desired global assessment of sensitivity over a range of dynamic responses of the system and not just at a single operating point. In this way, differential geometric samplers provide an efficient route from local to global sensitivity analysis of dynamical systems.

7

Discussion

7.1 Conclusions

Modern day science is becoming more and more dependent on the use of advanced statistical methodology, and this is being facilitated by more widely available high performance computing. A probabilistic Bayesian approach to statistical modelling is essential for characterising and reasoning with the inevitable uncertainty in the observations and measurements of complex systems in many different areas of the natural and physical sciences. We have seen in this thesis that the use of such a framework allows for the quantification of uncertainty both at a parameter level and a model level, although the development of efficient sampling methodology is essential in order for this framework to be useful in practice. Standard MCMC techniques allow us to draw samples from arbitrary probability distributions, however current methods perform poorly on models with large numbers of parameters and strong correlation structure.

In this thesis we have considered the use of Riemannian geometry in the context of developing novel MCMC sampling algorithms, with the Expected Fisher Information providing a natural link between statistics and differential geometry and allowing us to represent a statistical model as a Riemannian manifold.

We first reviewed existing MCMC approaches in Chapter 2, focusing on dynamical methods based on Langevin diffusions and Hamiltonian mechanics, and we noted that such methods are implicitly defined on a Euclidean or vector space. We then introduced the basic theory of differential geometry in Chapter 3 and presented a number of alternative sampling algorithms based on both Langevin diffusions and Hamiltonian

dynamics defined on a Riemannian manifold. We saw how the use of differential geometry has the potential to dramatically improve sampling efficiency, and we highlighted the link between local sensitivity analysis and the geometric structure induced by the Expected Fisher Information.

We investigated the efficiency of these new algorithms in Chapter 4 by performing simulation studies on a variety of challenging and topical statistical models, and we found that a pragmatic choice is to employ MCMC methods involving just the 2nd order geometric information, since they compromise between computational cost and effective sampling guided by the local geometry. Having derived and evaluated differential geometric MCMC methods on logistic regression models, stochastic volatility models, and log-Gaussian Cox models, we returned to the motivating problem of ODE modelling.

In Chapter 5 we derived the equations to allow us to employ differential geometric MCMC on models described by systems of nonlinear differential equations and we evaluated the performance of these algorithms, finding again that methods based assuming a locally constant metric gave the best time-normalised performance. We then proposed an approximate inference method for ordinary and delay differential equation models, and found that although this approach is computationally less expensive, especially for DDEs, it does not allow for marginal likelihoods to be calculated since a surrogate, approximate likelihood is employed. We therefore subsequently focused on inferential methods that explicitly solve the system of ODEs at each iteration. Finally we considered the challenges associated with performing a Bayesian analysis of ODE models using simple examples describing the spread of infection in a healthy population.

In the final chapter we returned to the original motivating example of modelling circadian rhythms using a larger system of nonlinear ordinary differential equations, demonstrating not only that we can perform inference on a model with unidentifiable parameters, but also obtain low variance estimates of the marginal likelihood, which we used to rank two competing model hypotheses; one with a negative feedback loop, and another with a positive feedback loop. We also investigated a cell signalling model that was based on commonly used mass-action kinetics and exhibited a different type of dynamic behaviour. We concluded that the use of differential geometry in MCMC allows us to use the local sensitivity information at each point to effectively perform a global sensitivity analysis of dynamical systems within a Bayesian framework.

In Appendix A we offer a summary of the manifold MCMC methods developed in this thesis, along with detailed pseudocode and guidelines for their application.

7.2 Future Work and Extensions

The work in this thesis presents a huge number of opportunities for further research, many of which have been brought up in discussion, documented in (77). The use of differential geometry as a framework for developing Monte Carlo algorithms allows us to exploit the wide variety of geometric ideas, structures and results that have been developed over the past 100 years or more. We could employ alternative metrics, based on more general divergence measures, or indeed alternative connections, which might turn out to be computationally more efficient for particular classes of statistical models. We could even consider non-Riemannian geometries, such as preferred point geometry, and such approaches may well be necessary to extend differential geometric MCMC to other types of models, for example those that currently require reversible jump methods for sampling multiple spaces of varying dimensionality. There is also without doubt much theoretical work still to be done, particularly for investigating the convergence and robustness properties of differential geometric MCMC methods.

From a practical perspective, there is much scope for improving the computational cost of this class of MCMC methodology. While the time normalised performance of these algorithms outperforms standard methods for many models, there is still plenty of room for further improving the computational scaling, which can be $\mathcal{O}(N^3)$ as a worst case scenario for simplified mMALA. Similarly, the main bottleneck for RMHMC is the generalised Leapfrog integration scheme, which requires the use of fixed point iterations to solve the implicitly defined equations of the Hamiltonian system, and so it would perhaps be worthwhile investigating alternative integration schemes for simulating these dynamics.

As we have seen, differential equation models can be particularly expensive to solve within an MCMC framework, although the careful use of parallel algorithms can help mitigate this cost. An alternative approach worth considering is the use of polynomial chaos expansions (76, 206), which can provide finite approximations to probability distributions and stochastic processes in terms of a set of basis functions. Such approaches are being more widely used in the field of engineering, however they have not yet had

a significant impact in the statistics community. It would be interesting to consider whether they might provide a computationally less expensive method of inference for the differential equation models considered in this thesis.

Finally, there are other modelling formalisms that are likely to benefit from a differential geometric perspective, such as partial and stochastic differential equations, indeed recent work on calculating the Fisher Information for SDEs (112) may well provide the natural link to allow the development of more efficient MCMC methods in this setting.

Appendix A

Manifold MCMC Recipes

In this appendix we briefly review the manifold MCMC methodology introduced in this thesis, detailing each algorithm and giving some guidance as to when each particular method may be most appropriately employed. All of these methods make use of the higher order geometric information available by considering the Riemannian manifold induced by the statistical model of interest. The metric tensor $G(\boldsymbol{\theta})$ may be given for example by the Expected Fisher Information at each point $\boldsymbol{\theta}$ in the parameter space.

A.1 Simplified Manifold MALA (SmMALA)

This algorithm is based on a simplification of the full mMALA algorithm and assumes a locally constant metric tensor at each point in the parameter space. The Gaussian proposal mechanism is defined with the mean given by the natural gradient and the covariance given by the inverse of the metric tensor. Pseudocode is given in Algorithm 4.

- Gives very good results for a wide range of models
- Quick and easy to code
- Requires only one metric tensor to be calculated per iteration, regardless of the model
- Particularly appropriate when the metric tensor and its derivatives are expensive to compute, e.g. models based on systems of ordinary differential equations

Algorithm 4 Simplified Manifold MALA

```

1: Initialise current  $\theta$ 
2: for IterationNum = 1 to NumSamples do
3:   Sample  $\theta^{\text{new}} \sim p(\theta^{\text{new}}|\theta) = \mathcal{N}(\mu(\theta, \epsilon), \Sigma(\theta, \epsilon))$ ,
      where  $\mu(\theta, \epsilon) = \theta + \frac{\epsilon^2}{2} G^{-1}(\theta) \nabla_{\theta} \mathcal{L}(\theta)$  and  $\Sigma(\theta, \epsilon) = \epsilon^2 G^{-1}(\theta)$ 
4:   Calculate current log-likelihood  $\mathcal{L}(\theta)$  and proposed log-likelihood  $\mathcal{L}(\theta^{\text{new}})$ 
5:   Calculate  $\log(p(\theta^{\text{new}}|\theta))$ ,  $\log(p(\theta|\theta^{\text{new}}))$ ,  $\log(\text{Prior}(\theta))$ ,  $\log(\text{Prior}(\theta^{\text{new}}))$ 
6:   LogRatio =  $\mathcal{L}(\theta^{\text{new}}) + \log(\text{Prior}(\theta^{\text{new}})) + \log(p(\theta|\theta^{\text{new}})) - \mathcal{L}(\theta) - \log(\text{Prior}(\theta)) - \log(p(\theta^{\text{new}}|\theta))$ 
7:   if LogRatio > 0 or LogRatio > log(rand) then
8:     Set  $\theta = \theta^{\text{new}}$ 
9:   end if
10: end for

```

A.2 Manifold MALA (mMALA)

The mMALA algorithm employs the Laplace-Beltrami operator to define the mean and covariance of a Gaussian proposal mechanism. This method takes into account the rate of change of the local coordinate system on the Riemannian manifold, i.e. the rate of change of the metric tensor. Pseudocode is given in Algorithm 5.

- Gives slightly better results than simplified mMALA in general
- Slightly more coding is required than simplified mMALA
- Requires one metric tensor and its derivatives with respect to each of the model parameters to be calculated per iteration
- Can give good results if the derivatives of the metric tensor are not too expensive to compute

A.3 Riemannian Manifold Hamiltonian Monte Carlo (RMHMC)

The RMHMC algorithm is the generalisation of Hamiltonian Monte Carlo to a Riemannian manifold. Instead of using the Leapfrog integrator, as in HMC, we must now

A.3 Riemannian Manifold Hamiltonian Monte Carlo (RMHMC)

Algorithm 5 Manifold MALA

- 1: Initialise current θ
 - 2: **for** IterationNum = 1 to NumSamples **do**
 - 3: Sample $\theta^{\text{new}} \sim p(\theta^{\text{new}}|\theta) = \mathcal{N}(\mu(\theta, \epsilon), \Sigma(\theta, \epsilon))$,
 where $\mu(\theta, \epsilon)_i = \theta_i + \frac{\epsilon^2}{2}(G^{-1}(\theta)\nabla_{\theta}\mathcal{L}(\theta))_i - \epsilon^2 \sum_{j=1}^D \left(G^{-1}(\theta) \frac{\partial G(\theta)}{\partial \theta_j} G^{-1}(\theta) \right)_{ij} +$
 $\frac{\epsilon^2}{2} \sum_{j=1}^D (G^{-1}(\theta))_{ij} \text{Tr} \left(G^{-1}(\theta) \frac{\partial G(\theta)}{\partial \theta_j} \right)$ and $\Sigma(\theta, \epsilon) = \epsilon^2 G^{-1}(\theta)$
 - 4: Calculate current log-likelihood $\mathcal{L}(\theta)$ and proposed log-likelihood $\mathcal{L}(\theta^{\text{new}})$
 - 5: Calculate $\log(p(\theta^{\text{new}}|\theta))$, $\log(p(\theta|\theta^{\text{new}}))$, $\log(\text{Prior}(\theta))$, $\log(\text{Prior}(\theta^{\text{new}}))$
 - 6: LogRatio = $\mathcal{L}(\theta^{\text{new}}) + \log(\text{Prior}(\theta^{\text{new}})) + \log(p(\theta|\theta^{\text{new}})) - \mathcal{L}(\theta) - \log(\text{Prior}(\theta)) -$
 $\log(p(\theta^{\text{new}}|\theta))$
 - 7: **if** LogRatio > 0 **or** LogRatio > log(rand) **then**
 - 8: Set $\theta = \theta^{\text{new}}$
 - 9: **end if**
 - 10: **end for**
-

use the generalised Leapfrog integrator to account for the fact that our Hamiltonian is non-separable. In this integration scheme, two of the equations are now defined implicitly, which we must solve using fixed point iterations. We note that when the statistical model induces a constant metric tensor, then the metric is independent of the parameters and the Hamiltonian becomes separable. In this case the generalised Leapfrog integrator reduces to the standard Leapfrog method and no longer requires the use of fixed point iterations. Pseudocode is given in Algorithm 6.

- High acceptance rate (close to 100%) and proposals are far from the current position
- Fixed point iterations may be slow if the metric tensor and its derivatives are expensive to compute
- Very fast when the statistical model induces a flat manifold, i.e. a manifold whose metric tensor is constant, e.g. the latent variables of the log-Gaussian Cox process in Chapter 4

A.4 Fixed Metric RMHMC

A pragmatic approach to obtaining excellent results using RMHMC is to employ a constant metric tensor obtained from some set of parameters from a region of high probability. For example, we could optimise (using the natural gradient) over the posterior and use the maximum a posteriori (MAP) parameter values. This approach is based on the assumption that the geometry of the statistical model does not change much in the high probability region of the posterior. The generalised Leapfrog algorithm then reduces to the standard Leapfrog integrator. Pseudocode is given in Algorithm 7.

- For most unimodal posterior distributions this approach will work well
- Very fast samples drawn using the standard Leapfrog method, while making use of the geometric information from the Riemannian manifold using the MAP parameter estimate
- High acceptance rate (close to 100%) and proposals are far from the current position
- This approach may not fare so well for posterior distributions with multiple modes that have vastly different local covariance structure

Algorithm 6 RMHMC with Generalised Leapfrog

```

1: Initialise current  $\theta$ 
2: for IterationNum = 1 to NumSamples do
3:   Sample new momentum  $\mathbf{p}^1 \sim p(\mathbf{p}^1|\theta) = \mathcal{N}(\mathbf{0}, G(\theta))$ 
4:   Calculate current  $H(\theta, \mathbf{p}^1)$ 
5:   Randomise  $N$  (leapfrog steps)
6:    $\theta^1 = \text{Current } \theta$ 
7:   for  $n = 1$  to  $N$  (leapfrog steps) do
8:     % Update the momentum with fixed point iterations
9:      $\hat{\mathbf{p}}^0 = \mathbf{p}^n$ 
10:    for  $i = 1$  to NumOfFixedPointSteps do
11:       $\hat{\mathbf{p}}^i = \mathbf{p}^n - \frac{\epsilon}{2} \nabla_{\theta} H(\theta^n, \hat{\mathbf{p}}^{i-1})$ 
      where  $\nabla_{\theta_i} H = -\frac{\partial \mathcal{L}(\theta)}{\partial \theta_i} + \frac{1}{2} \text{Tr} \left[ G(\theta)^{-1} \frac{\partial G(\theta)}{\partial \theta_i} \right] - \frac{1}{2} \mathbf{p}^T G(\theta)^{-1} \frac{\partial G(\theta)}{\partial \theta_i} G(\theta)^{-1} \mathbf{p}$ 
12:    end for
13:     $\mathbf{p}^{n+\frac{1}{2}} = \hat{\mathbf{p}}^i$ 
14:    % Update the parameters with fixed point iterations
15:     $\hat{\theta}^0 = \theta^n$ 
16:    for  $i = 1$  to NumOfFixedPointSteps do
17:       $\hat{\theta}^i = \theta^n + \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\theta^n, \mathbf{p}^{n+\frac{1}{2}}) + \frac{\epsilon}{2} \nabla_{\mathbf{p}} H(\hat{\theta}^{i-1}, \mathbf{p}^{n+\frac{1}{2}})$ 
      where  $\nabla_{\mathbf{p}} H(\theta, \mathbf{p}) = (G(\theta)^{-1} \mathbf{p})$ 
18:    end for
19:     $\theta^{n+1} = \hat{\theta}^i$ 
20:    % Update the momentum exactly
21:     $\mathbf{p}^{n+1} = \mathbf{p}^{n+\frac{1}{2}} - \frac{\epsilon}{2} \nabla_{\theta} H(\theta^{n+1}, \mathbf{p}^{n+\frac{1}{2}})$ 
    where  $\nabla_{\theta_i} H = -\frac{\partial \mathcal{L}(\theta)}{\partial \theta_i} + \frac{1}{2} \text{Tr} \left[ G(\theta)^{-1} \frac{\partial G(\theta)}{\partial \theta_i} \right] - \frac{1}{2} \mathbf{p}^T G(\theta)^{-1} \frac{\partial G(\theta)}{\partial \theta_i} G(\theta)^{-1} \mathbf{p}$ 
22:  end for
23:  Calculate proposed  $H(\theta^N, \mathbf{p}^N)$ 
24:  LogRatio =  $-\log(\text{Proposed } H) + \log(\text{Current } H)$ 
25:  % Accept or reject according to Metropolis ratio
26:  if LogRatio > 0 or LogRatio > log(rand) then
27:    Set  $\theta = \theta^N$ 
28:  end if
29: end for

```

Algorithm 7 Fixed Metric RMHMC with Standard Leapfrog

```

1: Optimise to obtain  $\theta_{MAP}$ 
2: Initialise current  $\theta = \theta_{MAP}$ 
3: for IterationNum = 1 to NumSamples do
4:   Sample new momentum  $\mathbf{p}^1 \sim p(\mathbf{p}^1 | \theta_{MAP}) = \mathcal{N}(\mathbf{0}, G(\theta_{MAP}))$ 
5:   Calculate current  $H(\theta, \mathbf{p}^1)$ 
6:   Randomise  $N$  (leapfrog steps)
7:    $\theta^1 = \text{Current } \theta$ 
8:   for  $n = 1$  to  $N$  (leapfrog steps) do
9:     % Update the momentum
10:     $\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^n - \frac{\epsilon}{2} \nabla_{\theta} H(\theta^n, \mathbf{p}^n)$ 
       where  $\nabla_{\theta_i} H = -\frac{\partial \mathcal{L}(\theta)}{\partial \theta_i}$ 
11:    % Update the parameters
12:     $\theta^{n+1} = \theta^n + \epsilon \nabla_{\mathbf{p}} H(\theta^n, \mathbf{p}^{n+\frac{1}{2}})$ 
       where  $\nabla_{\mathbf{p}} H(\theta, \mathbf{p}) = (G(\theta)^{-1} \mathbf{p})$ 
13:    % Update the momentum
14:     $\mathbf{p}^{n+1} = \mathbf{p}^{n+\frac{1}{2}} - \frac{\epsilon}{2} \nabla_{\theta} H(\theta^{n+1}, \mathbf{p}^{n+\frac{1}{2}})$ 
       where  $\nabla_{\theta_i} H = -\frac{\partial \mathcal{L}(\theta)}{\partial \theta_i}$ 
15:   end for
16:   Calculate proposed  $H(\theta^N, \mathbf{p}^N)$ 
17:   LogRatio =  $-\log(\text{Proposed } H) + \log(\text{Current } H)$ 
18:   % Accept or reject according to Metropolis ratio
19:   if LogRatio > 0 or LogRatio > log(rand) then
20:     Set  $\theta = \theta^N$ 
21:   end if
22: end for

```

Bibliography

- [1] L. ABBOTT. **Theoretical Neuroscience Rising**. *Neuron*, **60**(3):489–495, 2008. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0896627308008921>. 3
- [2] D. ALABADI, T. OYAMA, M. J. YANOVSKY, F. G. HARMON, P. MAS, AND S. A. KAY. **Reciprocal Regulation Between TOC1 and LHY/CCA1 within the Arabidopsis Circadian Clock**. *Science*, **293**(5531):880–883, 2001. Available from: <http://www.sciencemag.org/content/293/5531/880.short>. 163
- [3] M. P. ALLEN AND D. J. TILDESLEY. **Computer Simulation of Liquids**. Clarendon Press, 2006. Available from: <http://books.google.com/books?id=MtyCkgAACAAJ&printsec=frontcover>. 28
- [4] M. ALVAREZ, D. LUENGO, AND N. LAWRENCE. **Latent Force Models**. *Proceedings of The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, **5**:9–16, 2009. 128
- [5] S. AMARI. **Differential geometry of curved exponential families-curvatures and information loss**. *The Annals of Statistics*, **10**(2):357–385, 1982. Available from: <http://www.jstor.org/stable/2240672>. 69, 70
- [6] S. AMARI. **Natural gradient works efficiently in learning**. *Neural Computation*, **10**(2):251–276, 1998. Available from: <http://www.mitpressjournals.org/doi/abs/10.1162/089976698300017746>. 50, 51, 71
- [7] S. AMARI. **Divergence, Optimization and Geometry**. *ICONIP, Lecture Notes in Computer Science*, **5863**:185–193, 2009. Available from: <http://www.springerlink.com/index/OP6283075KVJ0708.pdf>. 64
- [8] S. AMARI AND H. NAGAOKA. **Methods of Information Geometry, Translations of Mathematical Monographs, vol. 191**. Oxford University Press, 2000. Available from: <http://www.lavoisier.fr/notice/frCWOR2AOYSRH3SO.html>. 48, 61, 62, 70
- [9] V. I. ARNOLD. **Ordinary Differential Equations**. The MIT Press, 1978. 24

- [10] C. ATKINSON AND A. MITCHELL. **Rao's Distance Measure.** *Sankhya: The Indian Journal of Statistics*, **43**(3):345–365, 1981. Available from: <http://www.jstor.org/stable/25050283>. 61
- [11] Y. BARD. **Nonlinear Parameter Estimation.** *Academic Press. New York*, 1974. Available from: <http://orton.catie.ac.cr/cgi-bin/wxis.exe/?IsisScript=CATALCO.xis&method=post&formato=2&cantidad=1&expresion=mfn=007559>. 114
- [12] M. BARENCO, D. TOMESCU, D. BREWER, R. CALLARD, J. STARK, AND M. HUBANK. **Ranked prediction of p53 targets using hidden variable dynamic modeling.** *Genome Biology*, **7**(3):25, 2006. Available from: <http://genomebiology.com/2006/7/3/R25>. xiii, 136, 137
- [13] A. BARKER. **Monte Carlo calculations of the radial distribution functions for a proton-electron plasma.** *Australian Journal of Physics*, **18**:119–133, 1965. Available from: <http://adsabs.harvard.edu/full/1965AuJPh..18..119B>. 13
- [14] O. E. BARNDORFF-NIELSEN. **Differential and Integral Geometry in Statistical Inference.** in *Differential Geometry in Statistical Inference*, *Institute of Mathematical Statistics*, 1987. 70
- [15] G. BATROUNI, G. KATZ, A. KRONFELD, G. LEPAGE, B. SVETITSKY, AND K. WILSON. **Langevin simulations of lattice field theories.** *Physical Review D*, **32**(10):2736–2747, 1985. Available from: http://prd.aps.org/abstract/PRD/v32/i10/p2736_1. 44
- [16] T. BAYES AND M. PRICE. **An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS.** *Philosophical Transactions (1683-1775)*, **53**:370–418, 1763. Available from: <http://www.jstor.org/stable/105741>. 9
- [17] V. BECKER, M. SCHILLING, J. BACHMANN, U. BAUMANN, A. RAUE, T. MAIWALD, J. TIMMER, AND U. KLINGMULLER. **Covering a broad dynamic range: information processing at the erythropoietin receptor.** *Science*, **328**(5984):1404–1408, 2010. Available from: <http://www.sciencemag.org/content/328/5984/1404.short>. 162, 165, 169, 185, 188
- [18] I. BEICHL AND F. SULLIVAN. **The Metropolis algorithm.** *Computing in Science and Engineering*, **2**(1):65–69, 2000. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=814660. 11
- [19] M. BENSON. **Parameter fitting in dynamic models.** *Ecological Modelling*, **6**(2):97–115, 1979. Available from: <http://linkinghub.elsevier.com/retrieve/pii/0304380079900292>. 114

- [20] G. BOOLE. **The calculus of logic**. *Cambridge and Dublin Mathematical Journal*, **3**:183–198, 1848. Available from: <http://www.maths.tcd.ie/pub/HistMath/People/Boole/CalcLogic/CalcLogic.pdf>. 8
- [21] G. E. P. BOX AND N. R. DRAPER. **Empirical model-building and response surfaces**. *Wiley*, 1987. Available from: <http://books.google.com/books?id=Q02dDRufJEAC&printsec=frontcover>. 156
- [22] K. BROWN AND J. SETHNA. **Statistical mechanical approaches to models with many poorly known parameters**. *Physical Review E*, **68**(2), 2003. Available from: <http://pre.aps.org/abstract/PRE/v68/i2/e021904>. 161
- [23] J. BURBEA AND C. R. RAO. **Entropy differential metric, distance and divergence measures in probability spaces: A unified approach**. *Journal of Multivariate Analysis*, **12**(4):575–596, 1982. Available from: <http://linkinghub.elsevier.com/retrieve/pii/0047259X82900653>. 64
- [24] J. BURBEA AND C. R. RAO. **On the convexity of some divergence measures based on entropy functions**. *IEEE Transactions on Information Theory*, **IT-28**(3):489–495, 1982. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1056497. 64
- [25] E. C. BUTCHER, E. L. BERG, AND E. J. KUNKEL. **Systems biology in drug discovery**. *Nature Biotechnology*, **22**(10):1253, 2004. Available from: <http://www.nature.com/nbt/journal/v22/n10/abs/nbt1017.html>. 160
- [26] B. CALDERHEAD. **A study of Population MCMC for estimating Bayes Factors over nonlinear ODE models**. *Masters Thesis, University of Glasgow*, 2008. Available from: <http://theses.gla.ac.uk/304/>. 6, 49, 81, 82
- [27] B. CALDERHEAD AND M. GIROLAMI. **Estimating Bayes factors via thermodynamic integration and population MCMC**. *Computational Statistics and Data Analysis*, (53):4028–4045, 2009. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0167947309002722>. xiv, 6, 14, 49, 81, 82, 120, 139, 144, 152, 160, 168, 180, 181
- [28] B. CALDERHEAD AND M. GIROLAMI. **Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods**. *Journal of the Royal Society Interface Focus*, **1**(6):821–835, 2011. Available from: <http://rsfs.royalsocietypublishing.org/content/1/6/821.short>. i, 158
- [29] B. CALDERHEAD, M. GIROLAMI, AND ND LAWRENCE. **Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes**. *Advances in Neural Information Processing Systems*, **21**:217–227, 2009. Available from: http://www.dcs.gla.ac.uk/publications/PAPERS/8980/GP_Inference_NIPS2008_FINAL.pdf. i, 115, 168

- [30] O. CALIN AND D. E. CHANG. **Geometric Mechanics on Riemannian Manifolds: Applications to Partial Differential Equations.** *Birkhauser*, 2005. Available from: <http://books.google.com/books?id=CQMqnMW9eMQC&printsec=frontcover>. 78
- [31] M. P. DO CARMO. **Differential Geometry of Curves and Surfaces.** *Prentice Hall*, 1976. Available from: <http://books.google.com/books?id=6BAZAQAAIAAJ&printsec=frontcover>. 54
- [32] M. P. DO CARMO. **Riemannian Geometry.** *Birkhauser Boston*, 1992. Available from: <http://books.google.com/books?id=ct91XCWkWEUC&printsec=frontcover>. 54, 57, 67, 69, 70, 161
- [33] N. N. CENCOV. **Statistical Decision Rules and Optimal Inference.** *Translations of Mathematical Monographs*, **53**, 1980. 62, 69
- [34] I. CHAVEL. **Riemannian Geometry: A Modern Introduction.** *Cambridge University Press*, 2006. Available from: http://books.google.com/books?id=3Gjp4vQ_mPkC&printsec=frontcover. 54
- [35] L. CHEN, Z. QIN, AND J. LIU. **Exploring Hybrid Monte Carlo in Bayesian Computation.** *Proceedings of International Society of Bayesian Analysis*, 2001. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.8023&rep=rep1&type=pdf>. 46
- [36] W. W. CHEN, M. NIEPPEL, AND P. K. SORGER. **Classic and contemporary approaches to modeling biochemical reactions.** *Genes and Development*, **24**(17):1861–1875, 2010. Available from: <http://genesdev.cshlp.org/content/24/17/1861.short>. 159
- [37] S. CHIB AND E. GREENBERG. **Understanding the Metropolis-Hastings algorithm.** *The American Statistician*, **49**(4):327–335, 1995. Available from: <http://www.jstor.org/stable/2684568>. 11
- [38] O. F. CHRISTENSEN, G. O. ROBERTS, AND J. S. ROSENTHAL. **Scaling limits for the transient phase of local Metropolis-Hastings algorithms.** *Journal of the Royal Statistical Society: Series B*, **67**(2):253–268, 2005. Available from: <http://eprints.lancs.ac.uk/19383/>. 44, 45, 101, 104, 106, 109
- [39] K. CHUNG. **Lectures from Markov processes to Brownian motion.** *New York: Springer*, 1982. 71
- [40] J. M. CORCUERA AND F. GIUMMOLE. **A Characterization of Monotone and Regular Divergences.** *Annals of the Institute of Statistical Mathematics*, **50**(3):433–450, 1998. Available from: <http://www.springerlink.com/index/R54J31U2W0563655.pdf>. 62

BIBLIOGRAPHY

- [41] R. COX. **Probability, frequency and reasonable expectation.** *American Journal of Physics*, **14**(1):1–13, 1946. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.164.4407&rep=rep1&type=pdf>. 2, 9
- [42] M. CREUTZ. **Global Monte Carlo algorithms for many-fermion systems.** *Physical Review D*, **38**(4):1228–1238, 1988. Available from: <http://link.aps.org/doi/10.1103/PhysRevD.38.1228>. 44
- [43] F. CRITCHLEY, P. MARRIOTT, AND M. SALMON. **Preferred point geometry and statistical manifolds.** *The Annals of Statistics*, **21**(3):1197–1224, 1993. Available from: <http://www.jstor.org/stable/2242195>. 70
- [44] F. CRITCHLEY, P. MARRIOTT, AND M. SALMON. **Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence.** *The Annals of Statistics*, **22**(3):1587–1602, 1994. Available from: <http://www.jstor.org/stable/2242241>. 70
- [45] I. CSISZAR. **A class of measures of informativity of observation channels.** *Period Math Hung*, **2**(1-4):191–213, 1972. Available from: <http://www.akademai.com/index/h45563437201516w.pdf>. 64
- [46] B. DANIELS, Y. CHEN, J. P. SETHNA, R. N. GUTENKUNST, AND C. R. MYERS. **Sloppiness, robustness, and evolvability in systems biology.** *Current Opinion in Biotechnology*, **19**(4):389–395, 2008. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0958166908000785>. 161
- [47] C. DARWIN. **The effects of cross and self fertilisation in the vegetable kingdom.** *London: John Murray*, 1876. 1
- [48] C. DARWIN AND F. DARWIN. **The Power of Movement in Plants.** *London: J. Murray*, 1880. 158
- [49] S. DAS, J. C. SPALL, AND R. GHANEM. **Efficient Monte Carlo computation of Fisher Information matrix using prior information.** *Computational Statistics and Data Analysis*, **54**(2):272–289, 2010. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4434830. 63
- [50] A. C. DAVISON. **Statistical models.** *Cambridge University Press*, page 726, 2003. Available from: <http://books.google.com/books?id=gQyIGGAiN4AC&printsec=frontcover>. 8
- [51] A. P. DAWID. **Discussion of Paper by B. Efron.** *The Annals of Statistics*, **3**:1231–1234, 1975. 69
- [52] A. P. DAWID. **Further Comments on a Paper by B. Efron.** *The Annals of Statistics*, **5**:1249, 1977. 69

- [53] C. T. J. DODSON AND T. POSTON. **Tensor geometry: the geometric viewpoint and its uses.** *Springer*, 1991. Available from: <http://books.google.com/books?id=PJjSPd70vpcC&printsec=frontcover>. 55
- [54] B. DOMSELAAR AND P. W. HEMKER. **Nonlinear parameter estimation in initial value problems.** *Stichting Mathematisch Centrum*, pages 1–49, 1975. Available from: <http://www.narcis.nl/publication/RecordID/oai:cwi.nl:9051>. 114
- [55] A. DOUCET, N. DE FREITAS, AND N. GORDON. **Sequential Monte Carlo methods in practice.** *Springer*, 2001. Available from: <http://books.google.com/books?id=uxX-koqKtMMC&printsec=frontcover>. 124
- [56] M. R. DOYLE, S. J. DAVIS, R. M. BASTOW, H. G. MCWATTERS, L. KOZMA-BOGNAR, F. NAGY, A. MILLAR, AND R. M. AMASINO. **The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana.** *Nature*, 419(6902):74, 2002. Available from: <http://www.nature.com/nature/journal/v419/n6902/abs/nature00954.html>. 163
- [57] S. DUANE, A. D. KENNEDY, B. J. PENDLETON, AND D. ROWETH. **Hybrid Monte Carlo.** *Physics Letters B*, 195(2):216–222, 1987. Available from: <http://linkinghub.elsevier.com/retrieve/pii/037026938791197X>. 17, 39, 44, 189
- [58] B. EFRON. **Defining the curvature of a statistical problem (with applications to second order efficiency).** *The Annals of Statistics*, 3(6):1189–1242, 1975. Available from: <http://www.jstor.org/stable/2958246>. 49, 59, 70
- [59] SHINTO EGUCHI. **Second Order Efficiency of Minimum Contrast Estimators in a Curved Exponential Family.** *The Annals of Statistics*, 11(3):793–803, 1983. Available from: <http://projecteuclid.org/euclid.aos/1176346246>. 70
- [60] S. EMMOTT. **Towards 2020 Science.** *Microsoft Research: Online Report*, 2005. Available from: http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/background_overview.htm. 3
- [61] K. ERGULER AND M. P. H. STUMPF. **Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models.** *Molecular BioSystems*, 7(5):1593–1602, 2011. Available from: <http://pubs.rsc.org/en/content/articlehtml/2011/mb/c0mb00107d>. 161
- [62] E. M. FARR AND S. KAY. **PRR7 protein levels are regulated by light and the circadian clock in Arabidopsis.** *The Plant Journal*, 52(3):548–560, 2007. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3113X.2007.03258.x/full>. 163
- [63] P. FERREIRA. **Extending Fisher’s measure of information.** *Biometrika*, 68(3):695–698, 1981. Available from: <http://www.jstor.org/stable/2335455>. 87

- [64] R. FLETCHER. **Practical Methods of Optimization**. Wiley, 2000. Available from: <http://books.google.com/books?id=LyHAQgAACAAJ&printsec=frontcover>. 66
- [65] D. FRENKEL AND B. SMIT. **Understanding molecular simulation: from algorithms to applications**. Academic Press, 2002. Available from: <http://books.google.com/books?id=5qTzldS9R0IC&printsec=frontcover>. 28, 29, 30
- [66] B. R. FRIEDEN. **Science from Fisher Information: A Unification**. Cambridge University Press, 2004. 62
- [67] N. FRIEL AND A. N. PETTITT. **Marginal likelihood estimation via power posteriors**. *Journal of the Royal Statistical Society: Series B*, **70**(3):589–607, 2008. Available from: <http://www3.interscience.wiley.com/journal/119418581/abstract>. 14, 81, 152, 168, 180, 188
- [68] D. GAMERMAN. **Sampling from the posterior distribution in generalized linear mixed models**. *Statistics and Computing*, **7**:57–68, 1997. Available from: <http://www.springerlink.com/index/L3738481750157T7.pdf>. 90
- [69] P. GAO, A. HONKELA, M. RATTRAY, AND N. LAWRENCE. **Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities**. *Bioinformatics*, **24**(16):70, 2008. Available from: <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/16/i70>. 136, 137
- [70] C. GEAR. **The automatic integration of ordinary differential equations**. *Communications of the ACM*, **14**(3):176–179, 1971. Available from: <http://portal.acm.org/citation.cfm?id=362566.362571>. 113
- [71] A. E. GELFAND AND A. F. M. SMITH. **Sampling-based approaches to calculating marginal densities**. *Journal of the American Statistical Association*, **85**(410), 1990. Available from: <http://www.jstor.org/stable/2289776>. 11
- [72] A. GELMAN, J. B. CARLIN, H. S. STERN, AND D. B. RUBIN. **Bayesian Data Analysis**. New York: Chapman and Hall, 2004. 87, 132
- [73] N. GERSHENFELD. **The Nature of Mathematical Modeling**. Cambridge University Press, 1999. Available from: <http://books.google.com/books?hl=en&lr=&id=1ST0h8U7NkkC&oi=fnd&pg=PR11&dq=the+nature+of+mathematical+modeling&ots=qitn3UJ8G9&sig=0H65IQyfmsE8y5nCT8wNePlf1ug>. 158
- [74] C. GEYER. **Practical Markov chain Monte Carlo**. *Statistical Science*, **7**(4):473–483, 1992. Available from: <http://www.jstor.org/stable/2246094>. 11, 86
- [75] C. J. GEYER. **Parallel Tempering**. In *Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface*, American Statistical Association, New York., page 156, 1991. 80

-
- [76] R. G. GHANEM AND P. D. SPANOS. **Stochastic finite elements: a spectral approach.** *Civil, Mechanical and Other Engineering Series*, 2003. Available from: <http://books.google.co.uk/books?id=WzgKyTQQcAwC>. 192
- [77] M. GIROLAMI AND B. CALDERHEAD. **Riemann manifold Langevin and Hamiltonian Monte Carlo methods.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(2):123–214, 2011. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.00765.x/full>. i, 13, 48, 85, 115, 161, 188, 192
- [78] T. GRAEPEL. **Solving noisy linear operator equations by Gaussian processes: Application to ordinary and partial differential equations.** *Proceedings of ICML*, 2003. Available from: <http://www.aaai.org/Papers/ICML/2003/ICML03-033.pdf>. 124
- [79] P. GREEN. **Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.** *Biometrika*, **82**(4):711–732, 1995. Available from: <http://biomet.oxfordjournals.org/content/82/4/711.short>. 14
- [80] U. GRENANDER AND M. I. MILLER. **Representations of knowledge in complex systems.** *Journal of the Royal Statistical Society: Series B*, **56**(4):549–603, 1994. Available from: <http://www.jstor.org/stable/2346184>. 44
- [81] R. GUTENKUNST, J. WATERFALL, F. P. CASEY, K. S. BROWN, C. R. MYERS, AND J. P. SETHNA. **Universally Sloppy Parameter Sensitivities in Systems Biology Models.** *PLoS Computational Biology*, **3**(10):189, 2007. Available from: <http://dx.plos.org/10.1371/journal.pcbi.0030189>. 161
- [82] H. HAARIO, M. LAINE, A. MIRA, AND E. SAKSMAN. **DRAM: Efficient adaptive MCMC.** *Statistics and Computing*, **16**(4):339–354, 2006. Available from: <http://www.springerlink.com/index/E1T2T818R3129T80.pdf>. 188
- [83] E. HAIRER, C. LUBICH, AND G. WANNER. **Geometric numerical integration: structure-preserving algorithms for ordinary differential equations.** *Berlin: Springer*, 2006. 19, 24, 25, 27, 39, 69, 79
- [84] A. HAJIAN. **Efficient cosmological parameter estimation with Hamiltonian Monte Carlo technique.** *Physical Review D*, **75**(8), 2007. Available from: <http://link.aps.org/doi/10.1103/PhysRevD.75.083525>. 43
- [85] W. HAMILTON. **On a General Method in Dynamics.** *Philosophical Transactions of the Royal Society*, pages 274–308, 1834. Available from: <http://www.emis.ams.org/classics/Hamilton/GenMeth.pdf>. 19
- [86] G. HAMMER, M. COOPER, F. TARDIEU, S. WELCH, B. WALSH, F. VAN EEUWIJK, S. CHAPMAN, AND D. PODLICH. **Models for navigating biological complexity in**

- breeding improved crop plants.** *Trends in Plant Science*, **11**(12):587–593, 2006. Available from: [http://linkinghub.elsevier.com/retrieve/pii/S1360-1385\(06\)00281-0](http://linkinghub.elsevier.com/retrieve/pii/S1360-1385(06)00281-0). 162
- [87] W. HASTINGS. **Monte Carlo sampling methods using Markov chains and their applications.** *Biometrika*, **57**(1):97–109, 1970. Available from: <http://biomet.oxfordjournals.org/content/57/1/97.short>. 11
- [88] J. HAVRDA AND F. CHARVT. **Quantification method of classification processes.** *Kybernetika*, **3**(1):30–35, 1967. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.683&rep=rep1&type=pdf>. 65
- [89] C. HIGHAM. **Bifurcation analysis informs Bayesian inference in the Hes1 feedback loop.** *BMC Systems Biology* 2009 3:12, **3**(1):12, 2009. Available from: <http://www.biomedcentral.com/1752-0509/3/12>. 170
- [90] A. C. HINDMARSH, P. N. BROWN, K. E. GRANT, S. L. LEE, R. SERBAN, D. E. SHUMAKER, AND C. S. WOODWARD. **SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers.** *ACM Transactions on Mathematical Software (TOMS)*, **31**(3):363–396, 2005. Available from: <http://portal.acm.org/citation.cfm?id=1089014.1089020>. 132, 169, 173
- [91] C. C. HOLMES AND L. HELD. **Bayesian auxiliary variable models for binary and multinomial regression.** *Bayesian Analysis*, **1**:145–168, 2005. Available from: http://www.stats.ox.ac.uk/~cholmes/Reports/Holmes_Held.pdf. 86, 89, 90
- [92] A. HONKELA, T. RAIKO, M. KUUSELA, M. TORNIO, AND J. KARHUNEN. **Approximate Riemannian Conjugate Gradient Learning for Fixed-Form Variational Bayes.** *The Journal of Machine Learning Research*, **11**:3235–3268, 2010. Available from: <http://portal.acm.org/citation.cfm?id=1953011.1953035>. 49
- [93] E. P. HSU. **Stochastic Analysis on Manifolds.** *Graduate Studies in Mathematics*, American Mathematical Society, 2002. Available from: <http://books.google.co.uk/books?id=2NM0Z7svRmEC>. 73
- [94] K. E. HUBBARD, F. C. ROBERTSON, N. DALCHAU, AND A. A. R. WEBB. **Systems Analyses of Circadian Networks.** *Molecular BioSystems*, **5**(12):1502–1511, 2009. Available from: <http://pubs.rsc.org/en/content/articlehtml/2009/mb/b907714f>. 162, 163
- [95] A. JAMES, J. MONREAL, G. NIMMO, C. KELLY, P. HERZYK, G. JENKINS, AND H. NIMMO. **The Circadian Clock in Arabidopsis Roots Is a Simplified Slave Version of the Clock in Shoots.** *Science*, **322**(5909):1832–1835, 2008. Available from: <http://www.sciencemag.org/content/322/5909/1832.short>. 163

-
- [96] A. JASRA, D. STEPHENS, AND C. HOLMES. **On population-based simulation for static inference.** *Statistics and Computing*, **17**:263–279, 2007. Available from: <http://www.springerlink.com/index/T15582622510271V.pdf>. 80, 130
- [97] E. T. JAYNES AND G. L. BRETTHORST. **Probability theory: the logic of science.** *Cambridge University Press*, 2003. Available from: <http://books.google.com/books?id=tTN4HuUNXjgC&printsec=frontcover>. 2, 9, 143
- [98] H. JEFFREYS. **Theory of Probability.** *1st ed. The Clarendon Press, Oxford*, 1939. 48
- [99] H. JEFFREYS. **An Invariant Form for the Prior Probability in Estimation Problems.** *Proceedings of the Royal Society of London: Series A*, **186**(1007):453–461, 1946. Available from: <http://www.jstor.org/stable/97883>. 61, 72
- [100] V. E. JOHNSON, S. G. KRANTZ, AND J. H. ALBERT. **Ordinal data modeling.** *New York: Springer*, 1999. Available from: <http://books.google.com/books?hl=en&lr=&id=6-Y8OL3sW14C&oi=fnd&pg=PA1&dq=Ordinal+Data+Modeling&ots=ka6F1AhF08&sig=Kms9dzJC1W2hY0ArxTBHmnrj4zI>. 90
- [101] R. E. KASS, B. CARLIN, A. GELMAN, AND R. M. NEAL. **Markov chain Monte Carlo in practice: A roundtable discussion.** *The American Statistician*, **52**(2):93–100, 1998. Available from: <http://www.questia.com/PM.qst?a=o&se=gglsc&d=5001349516>. 11
- [102] R. E. KASS AND A. E. RAFTERY. **Bayes factors.** *Journal of the American Statistical Association*, **90**(430):773–795, 1995. Available from: <http://www.jstor.org/stable/2291091>. 152
- [103] R. E. KASS AND P. W. VOS. **Geometrical Foundations of Asymptotic Inference.** *Wiley*, 1997. Available from: <http://books.google.com/books?id=7RzvAAAAAAJ&printsec=frontcover>. 49, 70
- [104] A. KENNEDY. **The Theory of Hybrid Stochastic Algorithms.** *NATO ASIB Proc. 224: Probabilistic Methods in Quantum Field Theory and Quantum Gravity*, pages 209–209, 1990. Available from: <http://adsabs.harvard.edu/abs/1990pmqf.conf..209K>. 44
- [105] J. KENT. **Time-reversible diffusions.** *Advances in Applied Probability*, **10**(4):819–835, 1978. Available from: <http://www.jstor.org/stable/1426661>. 43, 71, 73
- [106] S. KIM, N. SHEPHERD, AND S. CHIB. **Stochastic volatility: likelihood inference and comparison with ARCH models.** *Review of Economic studies*, **65**(3):361–393, 1998. Available from: <http://www3.interscience.wiley.com/journal/119112795/abstract>. 97

- [107] W. KIM, S. FUJIWARA, S. SUH, J. KIM, Y. KIM, L. HAN, K. DAVID, J. PUTTERILL, H. G. NAM, AND D. E. SOMERS. **ZEITLUPE is a circadian photoreceptor stabilized by GIGANTEA in blue light.** *Nature*, **449**(7160):356, 2007. Available from: <http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature06132.html>. 163
- [108] S. KIRKPATRICK. **Optimization by simulated annealing: Quantitative studies.** *Journal of Statistical Physics*, **34**(5-6):975–986, 1984. Available from: <http://www.springerlink.com/index/R8316332T1U15773.pdf>. 14
- [109] K. KNUTH AND J. SKILLING. **Foundations of Inference.** *Arxiv preprint arXiv:1008.4831*, 2010. Available from: <http://arxiv.org/abs/1008.4831>. 9
- [110] W. KOLCH, M. CALDER, AND D. GILBERT. **When kinases meet mathematics: the systems biology of MAPK signalling.** *FEBS Letters*, **579**(8):1891–1895, 2005. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0014579305001584>. 158
- [111] A. KOLMOGOROV. **Grundbegriffe der Wahrscheinlichkeitsrechnung.** *Springer, Berlin*, 1933. 2, 9
- [112] M. KOMOROWSKI, M. COSTA, D. A. RAND, AND M. STUMPF. **Sensitivity, robustness and identifiability in stochastic chemical kinetics models.** *Proceedings of the National Academy of Sciences*, **108**(21):8645–8650, 2011. Available from: <http://arxiv.org/abs/1104.1274>. 189, 193
- [113] E. KONUKOGLU, J. RELAN, U. CILINGIR, B. MENZE, P. CHINCHAPATNAM, A. JADIDI, H. COCHET, M. HOCINI, H. DELINGETTE, P. JAS, M. HASSAGUERRE, N. AYACHE, AND M. SERMESANT. **Efficient probabilistic model personalization integrating uncertainty on data and parameters: Application to Eikonal-Diffusion models in cardiac electrophysiology.** *Progress in Biophysics and Molecular Biology*, (Accepted), 2011. Available from: <http://www.sciencedirect.com/science/article/pii/S0079610711000654>. 159
- [114] M. KOORNNEEF AND D. MEINKE. **The development of Arabidopsis as a model plant.** *The Plant Journal*, **61**(6):909–921, 2010. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3113X.2009.04086.x/full>. 162
- [115] S. KULLBACK AND R. A. LEIBLER. **On Information and Sufficiency.** *The Annals of Mathematical Statistics*, **22**(1):79–86, 1951. Available from: <http://www.jstor.org/stable/2236703>. 64
- [116] P. LAMBERT AND P. EILERS. **Bayesian density estimation from grouped continuous data.** *Computational Statistics and Data Analysis*, **53**(4):1388–1399, 2009. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0167947308005616>. 45, 100

-
- [117] P. DE LAPLACE. **Theorie analytique des probabilités.** *M.V. Courcier*, 1820. Available from: <http://books.google.com/books?hl=en&lr=&id=cjAVAAAAQAAJ&oi=fnd&pg=PR1&dq=de+laplace&ots=TimpetdLEY&sig=ZWTEDqx1e3D5aLiobb3743G2gKY>. 9
 - [118] N. LARTILLOT AND H. PHILIPPE. **Computing Bayes factors using thermodynamic integration.** *Systematic Biology*, **55**(2):195–207, 2006. Available from: <http://sysbio.oxfordjournals.org/content/55/2/195.short>. 14, 81, 168, 180
 - [119] S. LAURITZEN. **Statistical manifolds (in Differential Geometry in Statistical Inference).** *Hayward: Institute of Mathematical Statistics*, pages 165–216, 1987. Available from: http://books.google.com/books?hl=en&lr=&id=idLwknoRcNIC&oi=fnd&pg=PA163&dq=statistical+manifolds&ots=23n47SNtlv&sig=51I3Y1Ds_HG9kIPOCDQYAm3t-Jw. 70
 - [120] N. LAWRENCE, M. SEEGER, AND R. HERBRICH. **Fast sparse Gaussian process methods: The informative vector machine.** *Advances in Neural Information Processing Systems 15*, 2003. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.4925&rep=rep1&type=pdf>. 130
 - [121] G. LEBANON. **Learning Riemannian Metrics.** *Proceedings of the 19th UAI*, 2003. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.3188&rep=rep1&type=pdf>. 49
 - [122] B. LEIMKUHLER AND S. REICH. **Simulating Hamiltonian Dynamics.** *Cambridge University Press*, 2004. Available from: <http://books.google.com/books?hl=en&lr=&id=tpb-tnsZi5YC&oi=fnd&pg=PR9&dq=simulating+hamiltonian+dynamics&ots=nKBg82kVlc&sig=psuwvNTwYTVb-ot-QE75MK1Aar4>. 23, 24, 33, 39
 - [123] F. LEVI, A. OKYAR, S. DULONG, P. F. INNOMINATO, AND J. CLAIRAMBAULT. **Circadian Timing in Cancer Treatments.** *Annual Review of Pharmacology and Toxicology*, **50**:377–421, 2010. Available from: <http://www-roc.inria.fr/bang/JC/LeviAnnuRevPharmacolToxicol2010.pdf>. 160
 - [124] C. VON LINNE AND S. FREER. **Linnaeus Philosophia Botanica.** *Oxford University Press*, page 402, 2005. Available from: <http://books.google.com/books?id=QstKWcHJyZgC&printsec=frontcover>. 2
 - [125] J. LIU. **Monte Carlo Strategies in Scientific Computing.** *New York: Springer*, 2008. Available from: <http://books.google.com/books?hl=en&lr=&id=R8E-yHaKCGUC&oi=fnd&pg=PR7&dq=monte+carlo+strategies+in+scientific+computing&ots=WbvXyPxb5K&sig=V37QHW8j8W2u4ZA261df7J4SaXk>. 11, 87, 96, 97, 100
 - [126] J. C. W. LOCKE, L. KOZMA-BOGNAR, P. D. GOULD, B. FEHER, E. KEVEI, F. NAGY, M. S. TURNER, A. HALL, AND A. MILLAR. **Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*.** *Molecular*

- Systems Biology*, **2**(1), 2006. Available from: <http://www.nature.com/msb/journal/v2/n1/full/msb4100102.html>. 159, 163
- [127] J. C. W. LOCKE, A. MILLAR, AND M. TURNER. **Modelling genetic networks with noisy and varied experimental data: the circadian clock in *Arabidopsis thaliana***. *Journal of Theoretical Biology*, **234**(3):383–393, 2005. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0022519304005879>. 159, 163
- [128] J. C. W. LOCKE, M. M. SOUTHERN, L. KOZMA-BOGNAR, V. HIBBERD, P. E. BROWN, M. S. TURNER, AND A. MILLAR. **Extension of a genetic network model by iterative experimentation and mathematical analysis**. *Molecular Systems Biology*, **1**(1), 2005. Available from: <http://www.nature.com/msb/journal/v1/n1/synopsis/msb4100018.html>. xiv, 159, 161, 163, 164, 165, 166, 169
- [129] J. D. LOGAN. **Applied Partial Differential Equations**. *Springer*, 2004. Available from: <http://books.google.com/books?id=zHvMVnzMYaMC&printsec=frontcover>. 5
- [130] P. MACKLIN, S. MCDUGALL, A. ANDERSON, M. CHAPLAIN, V. CRISTINI, AND J. LOWENGRUB. **Multiscale modelling and nonlinear simulation of vascular tumour growth**. *Journal of Mathematical Biology*, **58**(4-5):765–798, 2009. Available from: <http://www.springerlink.com/index/a0300580838767u7.pdf>. 159
- [131] P. MARRIOTT AND M. SALMON. **Applications of Differential Geometry to Econometrics**. *Cambridge University Press*, 2000. Available from: <http://books.google.com/books?id=1Jjm4I5tqkUC&printsec=frontcover>. 49, 54, 58
- [132] C. MCCLUNG. **Plant Circadian Rhythms**. *The Plant Cell*, **18**(4):792–803, 2006. Available from: <http://www.plantcell.org/content/18/4/792.short>. 158, 162, 163
- [133] N. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER. **Equation of state calculations by fast computing machines**. *The Journal of Chemical Physics*, **21**(6):1087–1092, 1953. Available from: <http://www.aliquote.org/pub/metropolis-et-al-1953.pdf>. 11
- [134] N. METROPOLIS AND S. ULAM. **The Monte Carlo method**. *Journal of the American Statistical Association*, **44**(247), 1949. Available from: <http://www.amstat.org/misc/TheMonteCarloMethod.pdf>. 10
- [135] D. MICHIE, D. J. SPIEGELHALTER, AND C. C. TAYLOR. **Machine Learning, Neural and Statistical Classification**. *Englewood Cliffs: Prentice Hall*, 1994. 89
- [136] W. MIO, D. BADLYANS, AND X. LIU. **A Computational Approach to Fisher Information Geometry with Applications to Image Analysis**. *Energy Minimization Methods in Computer Vision and Pattern Recognition: LNCS*, **3757**:18–33, 2005. 49

-
- [137] ANTONIETTA MIRA, REZA SOLGI, AND DANIELE IMPARATO. **Zero Variance Markov Chain Monte Carlo for Bayesian Estimators**. *arXiv ePrints*, Dec 2010. Available from: <http://arxiv.org/abs/1012.2983>. 18
 - [138] T. MIZUNO AND N. NAKAMICHI. **Pseudo-Response Regulators (PRRs) or True Oscillator Components (TOCs)**. *Plant and Cell Physiology*, **46**(5):677–685, 2005. Available from: <http://pcp.oxfordjournals.org/content/46/5/677.abstract>. 163
 - [139] L. MOHAMED, B. CALDERHEAD, M. FILIPPONE, M. CHRISTIE, AND M. GIROLAMI. **Population MCMC methods for history matching and uncertainty quantification**. *Computational Geosciences*, (1420-0597):1–14, 2011. Available from: <http://dx.doi.org/10.1007/s10596-011-9232-8>. 5, 159
 - [140] N. A. M. MONK. **Oscillatory Expression of Hes1, p53, and NF-B Driven by Transcriptional Time Delays**. *Current Biology*, **13**(16):1409–1413, 2003. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0960982203004949>. 134
 - [141] T. MULLER AND J. TIMMER. **Parameter identification techniques for partial differential equations**. *International Journal of Bifurcation and Chaos*, **14**(6):2053–2060, 2004. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.9933&rep=rep1&type=pdf>. 189
 - [142] P. MUNZ, I. HUDEA, J. IMAD, AND R. J. SMITH. **When zombies attack!: mathematical modelling of an outbreak of zombie infection**. *Infectious Disease Modelling Research Progress*, 2009. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.6699&rep=rep1&type=pdf>. 139, 147
 - [143] I. MURRAY, R. P. ADAMS, AND D. J. C. MACKAY. **Elliptical Slice Sampling**. *The Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, **9**:541–548, 2010. Available from: <http://arxiv.org/abs/1001.0175>. 14
 - [144] M. K. MURRAY AND J. W. RICE. **Differential Geometry and Statistics**. *New York: Chapman and Hall*, 1993. 58
 - [145] R. NEAL. **Probabilistic inference using Markov chain Monte Carlo methods**. *Technical Report: University of Toronto*, 1993. 40
 - [146] R. NEAL. **Bayesian learning for neural networks**. *New York: Springer*, 1996. Available from: http://books.google.com/books?hl=en&lr=&id=_peZjbrDC8cC&oi=fnd&pg=PR14&dq=bayesian+learning+for+neural+networks&ots=43H8VlXIBF&sig=qqofIyHcAvFN6ONavx-Cjt0l1Cg. 43, 46
 - [147] R. NEAL. **Slice Sampling**. *The Annals of Statistics*, **31**(3):705–741, 2003. Available from: <http://www.jstor.org/stable/3448413>. 14

- [148] R. NEAL. **MCMC using Hamiltonian dynamics**. *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press, 2010. Available from: <http://www.cs.utoronto.ca/~radford/ftp/ham-mcmc.ps>. 41, 42, 43
- [149] T. NOLAN, R. E. HANDS, AND S. A. BUSTIN. **Quantification of mRNA using real-time RT-PCR**. *Nature Protocols*, **1**(3):1559, 2006. Available from: <http://www.nature.com/nprot/journal/v1/n3/abs/nprot.2006.236.html>. 5
- [150] B. NOVAK AND J. J. TYSON. **Design Principles of Biochemical Oscillators**. *Nature Reviews Molecular Cell Biology*, **9**(12):981, 2008. Available from: <http://www.nature.com/nrm/journal/vaop/ncurrent/full/nrm2530.html>. 162
- [151] G. PARISI. **Correlation functions and computer simulations**. *Nuclear Physics B*, **180**(3):378–384, 1981. Available from: <http://linkinghub.elsevier.com/retrieve/pii/0550321381900560>. 43
- [152] D. H. PARK, D. E. SOMERS, Y. S. KIM, Y. H. CHOY, H. K. LIM, M. S. SOH, H. J. KIM, S. A. KAY, AND H. G. NAM. **Control of Circadian Rhythms and Photoperiodic Flowering by the Arabidopsis GIGANTEA Gene**. *Science*, **285**(5433):1579–1582, 1999. Available from: <http://www.sciencemag.org/content/285/5433/1579.short>. 163
- [153] X. PENNEC. **Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements**. *Journal of Mathematical Imaging and Vision*, **25**(127-154), 2006. Available from: <http://www.springerlink.com/index/7543M49347H2U1Q2.pdf>. 49
- [154] P. H. PESKUN. **Optimum Monte-Carlo sampling using Markov chains**. *Biometrika*, **60**(3):607–612, 1973. Available from: <http://dx.doi.org/10.1093/biomet/60.3.607>. 13
- [155] A. PETER AND A. RANGARAJAN. **A new closed-form information metric for shape analysis**. *Medical Image Computing and Computer-assisted Intervention LNCS*, **4190**:249–256, 2006. Available from: <http://www.springerlink.com/index/u4775u111743144p.pdf>. 66
- [156] K. B. PETERSEN AND M. S. PEDERSEN. **The Matrix Cookbook**. *Technical University of Denmark*, 2008. Available from: <http://matrixcookbook.com/>. 74
- [157] A. POYTON, M. VARZIRI, K. B. MCAULEY, P. J. MCLELLAN, AND J. O. RAMSAY. **Parameter estimation in continuous-time dynamic models using principal differential analysis**. *Computers and Chemical Engineering*, **30**:698–708, 2006. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0098135405003042>. 123

-
- [158] Y. QI AND T. MINKA. **Hessian-based Markov chain Monte Carlo algorithms.** *1st Cape Cod Workshop on Monte Carlo Methods*, 2002. Available from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.2754>. 66
 - [159] J. RAMSAY, G. HOOKER, D. CAMPBELL, AND J. CAO. **Parameter estimation for differential equations: a generalized smoothing approach.** *Journal of the Royal Statistical Society: Series B*, **69**(5):741–796, 2007. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2007.00610.x/full>. 119, 123, 124, 130, 131, 132, 133, 157
 - [160] J. O. RAMSAY AND B. W. SILVERMAN. **Functional data analysis.** *Springer*, 2005. Available from: http://books.google.com/books?id=mU3dop5wY_4C&printsec=frontcover. 123
 - [161] C. R. RAO. **Information and the accuracy attainable in the estimation of several parameters.** *Calcutta Mathematical Bulletin*, **37**:81–91, 1945. 48, 50, 61, 62
 - [162] D. RAPAPORT. **The Art of Molecular Dynamics Simulation.** *Cambridge University Press*, 2004. Available from: http://books.google.com/books?hl=en&lr=&id=iqDJ2hjQBM&oi=fnd&pg=PR9&dq=molecular+dynamics+perturbations&ots=krIMszfpXQ&sig=zvtrz0P-KlY5sDFWi0_tZtt_tek. 28
 - [163] C. RASMUSSEN. **Gaussian Processes to Speed up Hybrid Monte Carlo for Expensive Bayesian Integrals.** *Bayesian Statistics*, **7**:651–659, 2003. 124
 - [164] C. E. RASMUSSEN AND C. WILLIAMS. **Gaussian Processes for Machine Learning.** *Cambridge: MIT Press*, 2006. Available from: <http://books.google.com/books?id=vWtwQgAACAAJ&printsec=frontcover>. 126, 127
 - [165] A. RAUE, C. KREUTZ, T. MAIWALD, J. BACHMANN, M. SCHILLING, U. KLINGMULLER, AND J. TIMMER. **Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood.** *Bioinformatics*, **25**(15):1923–1929, 2009. Available from: <http://bioinformatics.oxfordjournals.org/content/25/15/1923.short>. 160, 161, 167, 189
 - [166] B. RIPLEY. **Pattern Recognition and Neural Networks.** *Cambridge University Press*, 2008. Available from: <http://books.google.com/books?hl=en&lr=&id=m12UR8QmLqoC&oi=fnd&pg=PR9&dq=pattern+recognition+and+neural+networks&ots=aLQpeLYGXd&sig=KRTqhilPk5BIi3BabEP47bH4FS8>. 89
 - [167] C. P. ROBERT AND G. CASELLA. **Monte Carlo Statistical Methods.** *New York: Springer*, 2004. Available from: http://books.google.com/books?hl=en&lr=&id=HfhGAXn5GugC&oi=fnd&pg=PR10&dq=monte+carlo+statistical+methods&ots=ByA0_UeWRz&sig=7GJvWz-v8XbugVS0ldZeT0iw8Fw. 10, 11, 12, 13, 86, 160

-
- [168] G. O. ROBERTS AND J. S. ROSENTHAL. **Markov chain Monte Carlo: some practical implications of theoretical results.** *Canadian Journal of Statistics*, **26**(1):5–20, 1998. Available from: <http://www3.interscience.wiley.com/journal/121577893/abstract>. 44, 86
- [169] G. O. ROBERTS AND J. S. ROSENTHAL. **Optimal scaling of discrete approximations to Langevin diffusions.** *Journal of the Royal Statistical Society: Series B*, **60**(1):255–268, 1998. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00123/abstract>. 44
- [170] G. O. ROBERTS AND J. S. ROSENTHAL. **Optimal scaling for various Metropolis-Hastings algorithms.** *Statistical Science*, **16**(4):351–367, 2001. Available from: <http://www.projecteuclid.org/GetRecord?id=euclid.ss/1015346320>. 18
- [171] G. O. ROBERTS AND R. L. TWEEDIE. **Exponential convergence of Langevin distributions and their discrete approximations.** *Bernoulli*, **2**(4):341–363, 1996. Available from: <http://www.jstor.org/stable/3318418>. 44
- [172] P. ROSSKY, J. D. DOLL, AND H. L. FRIEDMAN. **Brownian dynamics as smart Monte Carlo simulation.** *The Journal of Chemical Physics*, **69**(10):4628–4633, 1978. Available from: <http://link.aip.org/link/?JCPA6/69/4628/1>. 43
- [173] H. RUE AND L. HELD. **Gaussian Markov random fields: theory and applications.** *Monographs on statistics and applied probability*, Chapman and Hall, 2005. Available from: <http://books.google.co.uk/books?id=TLBYs-faw-0C>. 99
- [174] H. RUE, S. MARTINO, AND N. CHOPIN. **Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.** *Journal of the Royal Statistical Society: Series B*, **71**(2):319–392, 2009. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2008.00700.x/full>. 14, 111
- [175] A. SALTELLI, S. TARANTOLA, AND F. CAMPOLONGO. **Sensitivity analysis as an ingredient of modeling.** *Statistical Science*, pages 377–395, 2000. Available from: <http://www.jstor.org/stable/10.2307/2676831>. 6
- [176] RICHARD SAVAGE AND SEB OLIVER. **Bayesian Methods of Astronomical Source Extraction.** *The Astrophysical Journal*, **661**:1339, 2007. Available from: <http://iopscience.iop.org/0004-637X/661/2/1339>. 3
- [177] C. SAWYERS. **Targeted cancer therapy.** *Nature*, **432**(7015):294, 2004. Available from: <http://www.nature.com/nature/journal/v432/n7015/abs/nature03095.html>. 160
- [178] H. SCHMIDT AND M. JIRSTRAND. **Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology.** *Bioinformatics*, **22**(4):514–515, 2006. Available from: <http://bioinformatics.oxfordjournals.org/content/22/4/514.short>. 132, 169

- [179] H. SHANKARAN, H. RESAT, AND H. S. WILEY. **Cell Surface Receptors for Signal Transduction and Ligand Transport: A Design Principles Study.** *PLoS Computational Biology*, **3**(6):101, 2007. Available from: <http://dx.plos.org/10.1371/journal.pcbi.0030101>. 165
- [180] D. S. SIVIA AND J. SKILLING. **Data analysis: a Bayesian tutorial.** *Oxford University Press*, 2006. Available from: <http://books.google.com/books?id=zN-yliq6eZ4C&printsec=frontcover>. 143
- [181] J. SKILLING. **Bayesian solution of ordinary differential equations.** *Maximum entropy and Bayesian methods: proceedings of the Eleventh International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle*, 1992. 124
- [182] J. SKILLING. **Nested Sampling for General Bayesian Computation.** *Bayesian Analysis*, **1**(4):833–860, 2006. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.5542&rep=rep1&type=pdf>. 14
- [183] E. SNELSON, C. E. RASMUSSEN, AND Z. GHAHRAMANI. **Warped Gaussian Processes.** *Advances in Neural Information Processing Systems 16*, 2004. Available from: http://books.google.com/books?hl=en&lr=&id=0F-9C7K8fQ8C&oi=fnd&pg=PA337&dq=Warped+Gaussian+processes&ots=TFMzk_X76_&sig=rFGyMXBC5M01jJX0Uh6bvMg9eKU. 126
- [184] E. SOLAK, R. MURRAY-SMITH, W. E. LEITHEAD, D. J. LEITH, AND C. E. RASMUSSEN. **Derivative observations in Gaussian process models of dynamic systems.** *Advances in Neural Information Processing Systems 15*, 2003. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.5525&rep=rep1&type=pdf>. 127
- [185] C. SOMERVILLE AND M. KOORNNEEF. **A fortunate choice: the history of Arabidopsis as a model plant.** *Nature Reviews Genetics*, **3**(11):883, 2002. Available from: <http://www.nature.com/nrg/journal/v3/n11/abs/nrg927.html>. 162
- [186] J. SPALL. **Monte Carlo computation of the Fisher Information matrix in non-standard settings.** *Journal of Computational and Graphical Statistics*, **14**(4):889–909, 2005. Available from: <http://pubs.amstat.org/doi/abs/10.1198/106186005X78800.63>. 63
- [187] M. SPIVAK. **A Comprehensive Introduction to Differential Geometry.** *Publish or Perish*, 1979. Available from: <http://books.google.com/books?id=9ozgTgEACAAJ&printsec=frontcover>. 54
- [188] J. STELLING, E. D. GILLES, AND F. J. DOYLE. **From the Cover: Robustness properties of circadian clock architectures.** *Proceedings of the National Academy of*

- Sciences*, **101**(36):13210–13215, 2004. Available from: <http://www.pnas.org/content/101/36/13210.short>. 162
- [189] S. M. STIGLER. **Darwin, Galton and the Statistical Enlightenment.** *Journal of the Royal Statistical Society: Series A*, **173**(3):469–482, 2010. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2010.00643.x/full>. 2
- [190] G. SUSSMAN AND J. WISDOM. **Chaotic Evolution of the Solar System.** *Science*, **257**(5066):56–62, 1992. Available from: <http://www.sciencemag.org/content/257/5066/56.short>. 29
- [191] I. SWAMEYE, T. MULLER, J. TIMMER, O. SANDRA, AND U. KLINGMULLER. **Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling.** *Proceedings of the National Academy of Sciences*, **100**(3):1028–1033, 2003. Available from: <http://www.pnas.org/content/100/3/1028.short>. 188
- [192] J. S. TAKAHASHI, H. HONG, C. H. KO, AND E. L. MCDEARMON. **The genetics of mammalian circadian order and disorder: implications for physiology and disease.** *Nature Reviews Genetics*, **9**(10):764, 2008. Available from: <http://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg2430.html>. 159
- [193] A. TARANTOLA. **Inverse problem theory and methods for model parameter estimation.** *Society for Industrial and Applied Mathematics*, 2005. Available from: <http://books.google.com/books?hl=en&lr=&id=kEboSYSU-nAC&oi=fnd&pg=PR11&dq=Inverse+Problem+Theory+and+Methods+for+Model+Parameter+Estimation&ots=V01cisfwXq&sig=009y49KtyvchDasWbBo6yRBE89c>. 113
- [194] S. C. THAIN, A. HALL, AND A. MILLAR. **Functional independence of circadian clocks that regulate plant gene expression.** *Current Biology*, **10**(16):951–956, 2000. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0960982200006308>. 163
- [195] C. J. TOMLIN AND J. D. AXELROD. **Biology by numbers: mathematical modelling in developmental biology.** *Nature Reviews Genetics*, **8**(5):331, 2007. Available from: <http://www.nature.com/nrg/journal/v8/n5/abs/nrg2098.html>. 158
- [196] T. TONI, D. WELCH, N. STRELKOWA, A. IPSEN, AND M. P. H. STUMPF. **Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems.** *Journal of the Royal Society Interface*, **6**(31):187–202, 2009. Available from: <http://rsif.royalsocietypublishing.org/content/6/31/187.full>. 160, 188
- [197] M. TRANSTRUM, B. B. MACHTA, AND J. P. SETHNA. **Geometry of nonlinear least squares with applications to sloppy models and optimization.** *Physical Review*

- E*, **83**(3):036701, 2011. Available from: <http://pre.aps.org/abstract/PRE/v83/i3/e036701>. 161, 167, 170, 175, 189
- [198] R. TSUTAKAWA. **Design of experiment for bioassay**. *Journal of the American Statistical Association*, **67**(339):584–590, 1972. Available from: <http://www.jstor.org/stable/2284443>. 87
- [199] P. TURQ, F. LANTELME, AND H. L. FRIEDMAN. **Brownian dynamics: Its application to ionic solutions**. *The Journal of Chemical Physics*, **66**(7):3039, 1977. Available from: <http://link.aip.org/link/doi/10.1063/1.434317/html>. 43
- [200] M. VALLISNERI. **Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects**. *Physical Review D*, **77**(4):042001, 2008. Available from: <http://link.aps.org/doi/10.1103/PhysRevD.77.042001>. 167, 168, 170
- [201] J. VANHATALO AND A. VEHTARI. **Sparse log Gaussian processes via MCMC for spatial epidemiology**. *JMLR: Workshop and Conference Proceedings: Gaussian Processes in Practice*, **1**:73–89, 2007. Available from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.68.8443>. 111
- [202] J. M. VARAH. **A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations**. *SIAM Journal on Scientific and Statistical Computing*, **3**(1):28–46, 1982. Available from: <http://link.aip.org/link/doi/10.1137/0903003/html>. 114, 123, 130
- [203] V. VYSHEMIRSKY AND M. GIROLAMI. **Bayesian ranking of biochemical system models**. *Bioinformatics*, **24**(6):833–839, 2008. Available from: <http://bioinformatics.oxfordjournals.org/content/24/6/833.short>. 4, 124, 139, 159, 160
- [204] D. J. WILKINSON. **Stochastic Modelling for Systems Biology**. *Taylor and Francis*, 2006. Available from: <http://books.google.com/books?id=roHTk4m8JGAC&printsec=frontcover>. 5, 159, 189
- [205] T. WILLMORE. **Riemannian Geometry**. *Oxford University Press*, 1997. Available from: <http://books.google.com/books?id=J0nvAAAAAAAJ&printsec=frontcover>. 54
- [206] D. XIU. **Numerical Methods for Stochastic Computations: A Spectral Method Approach**. *Princeton University Press*, 2010. Available from: <http://books.google.co.uk/books?id=GY9qyJd4CvQC>. 192
- [207] T. XU, V. VYSHEMIRSKY, A. GORMAND, A. VON KRIEGSHEIM, M. GIROLAMI, G. S. BAILLIE, D. KETLEY, A. J. DUNLOP, G. MILLIGAN, M. D. HOUSLAY, AND W. KOLCH. **Inferring Signaling Pathway Topologies from Multiple Perturbation Measurements of Specific Biochemical Species**. *Science Signaling*, **3**(113):20, 2010. Available

BIBLIOGRAPHY

from: <http://stke.sciencemag.org/cgi/content/full/sigtrans;3/113/ra20>. 159, 160, 164, 188

- [208] M. ZLOCHIN AND Y. BARAM. **Manifold stochastic dynamics for Bayesian learning.** *Neural Computation*, **13**(11):2549–2572, 2001. Available from: <http://www.mitpressjournals.org/doi/abs/10.1162/089976601753196021>. 77